12-1994

High Breakdown Rank-Based Estimates for Linear Models

William H. Chang
Western Michigan University

HIGH BREAKDOWN RANK-BASED ESTIMATES FOR LINEAR MODELS

by

William H. Chang

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Mathematics and Statistics

Western Michigan University
Kalamazoo, Michigan
December 1994

HIGH BREAKDOWN RANK-BASED ESTIMATES
FOR LINEAR MODELS

William H. Chang, Ph.D.

Western Michigan University, 1994

In this dissertation we are concerned with the linear model $y_i = \underline{x}_i' \underline{\beta} + \epsilon_i$, $i = 1, \ldots, n$, where $\underline{x}_i$ is a p-dimensional vector of known constants, $\underline{\beta}$ is a p-dimensional vector of unknown parameters, and ϵ_i's are random errors. The least squares estimates fit the linear model quite well, when ϵ_i's are independent and identically distributed with $N(0, \sigma^2)$ for some $\sigma > 0$. Unfortunately, the least squares estimates may be spoiled by small but reasonable deviations from normal structure (Huber 1972, Andrews 1974).

In 1983, Sievers proposed the general rank estimates by minimizing the dispersion function $D(\underline{\beta}) = \sum_{i<j} b_{ij} \mid z_i - z_j \mid$, where weights b_{ij} are functions of $\underline{X}$. Naranjo and Hettmansperger (1990) proved that for appropriate weights, the influence function of the Sievers's general rank estimates is bounded in both the X and Y spaces, and the breakdown point $\epsilon^* \leq 1/3$. The estimates downweight outliers indiscriminately in the X space, even if the point may fit the model.

In this dissertation, we extend the work of Sievers (1983) by defining the weights as functions of both $\underline{X}$ and $\underline{Y}$. We derived the asymptotic distribution of the estimates. A high breakdown point of 50% for the estimates has also been proved.

Order Number 9517439

High-breakdown rank-based estimates for linear models

Chang, William Hsuming, Ph.D.

Western Michigan University, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

ACKNOWLEDGEMENTS

I wish to thank Dr. Joseph W. McKean for his advice and support during this research and during my years as a graduate student at Western Michigan University. I would also like to thank Dr. Joshua D. Naranjo for his help.

William Chang

TABLE OF CONTENTS

Table of Contents - Continued

LIST OF TABLES

LIST OF FIGURES

List of Figures - Continued

CHAPTER I

INTRODUCTION

Linear model is one of the most popular models used in statistical analysis and statistical experimental designs. In this thesis, we are concerned with the following linear model

$$y_i = \underline{x}_i' \underline{\beta} + \epsilon_i, \quad i = 1, \ldots, n,$$

where is $\underline{x}_i$ a p-dimensional vector of known regression constants, $\underline{\beta}$ is a p-dimensional vector of unknown parameters, and ϵ_i's are random errors. There are two standard assumptions for this model: ϵ_i's are independent and identical distributed with some F, and the F is normally distributed with $N(0, \sigma^2)$ for some $\sigma > 0$.

The least squares estimates, introduced by Gauss and Legendre (Plackett 1972, Stigler 1981), are the most popular estimates. It fits the linear model quite well, when the standard assumptions are held. Unfortunately, the least squares estimates may be spoiled by small but reasonable deviations from normal structure (Huber 1972, Andrews 1974).

Because real data usually do not completely satisfy the standard assumptions, the quality of the statistical analysis based on the least squares estimates can be dramatically affected (student 1927, Pearson 1931, Box 1953, and Tukey 1960). Hample, Ronchetti, Rousseeuw, and Stahel (1986) thought that routine data contains 1% to 10% gross errors.

Because the standard assumptions are frequently violated in the real world, the field of robust statistics has been developing rapidly in recent years. These robust estimates are generally highly efficient compared to the least squares estimates on data which meet the standard assumptions, but are much less sensitive than the least squares estimates for data with outliers.

In 1972, Jaechel proposed the regular rank estimates by minimizing the dispersion function defined as

$$D(\beta) = \sum_{i=1}^{n} a(R(y_i - \underset{\sim}{x}_i' \underset{\sim}{\beta}))(y_i - \underset{\sim}{x}_i' \underset{\sim}{\beta}),$$

where $R(u_i)$ denotes the rank of u_i among $u_1, u_2, \ldots, u_n$ and $a(1) \leq a(2) \leq \ldots \leq a(n)$ is a set of scores such that

$$\sum_{i=1}^{n} a(i) = 0 .$$

McKean and Hettmansperger(1976), (1977), (1978), Hettmansperger(1984), McKean and Sievers(1989), McKean et al.(1989) further extended Jaechel's approach. These estimates are more robust than the least squares estimates. The influence function of these estimates are bounded in Y space, but they are still sensitive to outliers in X space.

In 1983, Sievers proposed the general rank estimates which are obtained by minimizing the dispersion function defined as:

$$D(\beta) = \sum_{i=1}^{n} b_{ij} \mid z_j - z_i \mid ,$$

where b_{ij} are weights which are functions of $\underline{X}$, and residual z_j is defined as

$$z_i = y_i - \underline{x}_i' \underline{\beta} .$$

In 1990, Naranjo and Hettmansperger proved that for appropriate weights, the influence function of the Sievers's general rank estimates is bounded in both the X and Y spaces. And the breakdown point of the Sievers's general rank estimates $\epsilon^* \leq 1/3$. The Sievers's estimates downweight outliers indiscriminately in X space, even if some outliers may fit the model.

In general, we can divided the weight functions into three groups. The first group consists of constant weight functions. Specially, when weight functions are equal, the dispersion function is equivalent to Jaeckel's dispersion function with Wilcoxon scores (Hettmansperger and McKean, 1978). With such equal weights, the analysis and calculation of the coefficients can be greatly simplified. Because weight functions are constants, the outliers in X space are not able to be downweighted, and the influence function of these estimates are not bounded in X space.

In the second group, the weights are functions of $\underline{X}$. With these weight functions, we are able to downweight high leverage points. the weights can be chosen so that their influence functions are bounded in both the X and Y spaces. Therefore, the estimates obtained by models with these weight functions are more robust than ones from the first group.

The last group consists of the weights which are functions of both $\underline{X}$ and $\underline{Y}$. The weight functions which will be used in this paper belong to this group. Because some high leverage points could be good points, using these weight functions we will be able to only downweight the bad leverage points and keep the good leverage points. Such weight functions provide us a possibility of obtaining robust estimates with high efficiency.

Krasker Welsch(1982) stated: "Any down weighing in X space that does not include some consideration for how the y values at these outlying observations fit the pattern set by the bulk of the data cannot be efficient." In this paper, we will extend the work of Sievers(1983) by defining the weights as a function of both $\underline{X}$ and $\underline{Y}$. The goal that we seek to achieve was stated by Yohai and Zamar(1988): One of the goals of the robust regression estimation is to achieve simultaneously the following three properties: a breakdown point of roughly 50%, a bounded influence function, and a high efficiency vs. the least squares estimates when F is Gaussian.

In the following chapters, we will discuss the estimates obtained by minimizing the dispersion function with the weights which are functions of both $\underline{X}$ and $\underline{Y}$. We will derive the asymptotic distribution of the estimates. A high breakdown point of 50% for the estimates will also be proved.

CHAPTER II

ESTIMATION OF THE REGRESSION COEFFICIENTS

2.1 Notation

In this thesis, we will discuss the linear model, the estimates of the regression coefficients, and some robust properties associated with the estimates. In order to derive these robust properties, we first need to state some notations and assumptions which will be used through this thesis. The linear model that we will use can be expressed as:

$$y_i = \underline{x}_i' \, \underline{\beta} + \epsilon_i , \quad i = 1, \ldots , n,$$

where y_i is an observation, $\underline{x}_i$ is a p-dimensional vector of known regression constants, and $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters. ϵ_i , $\epsilon_2 \ldots$, ϵ_n are independent and identically distributed random variables with density f.

Let $\underline{X}$ be a $n \times p$ design matrix which is defined as

$$\underline{X} = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_n' \end{bmatrix} .$$

Let $\underline{Y}$ be an $n \times 1$ vector of observations and $\underline{E}$ be an $n \times 1$ vector of random variables which are defined as:

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad \text{and} \quad \underline{E} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}.$$

Then, we can express the above linear model in the matrix form:

$$\underline{Y} = \underline{X}\,\underline{\beta} + \underline{E}.$$

As we proceed, we will make further assumptions about the design matrix $\underline{X}$ and the density f of random variables. If desired, we can further define the random variable ϵ_i as:

$$\epsilon_i = \alpha + \tau_i, \quad i = 1, \dots, n,$$

where α is an intercept parameter. In this way, we can take $\underline{X}$ to be a centered design matrix with out loss of generality. This will be proven convenient later.

2.2 The Geometry of the Estimation Procedure

Once having the linear model, the next task that we are facing is to find a method to estimate the vector of the regression coefficients $\underline{\beta}$. In this thesis, we will estimate $\underline{\beta}$ by minimizing the dispersion function defined as:

$$D(\beta) = \sum_{i<j} b_{ij} \mid z_j - z_i \mid ,$$

where z_i is the residual which is defined as $z_i = y_i - \underline{x}_i' \underline{\beta}$, and b_{ij} are the weights defined as follows:

Let $\psi(t) = 1,\ t,\ -1,$ as $t > 1,\ -1 \le t \le 1,\ t < 1$ and define the weights

$$b_{ij} = \psi\left[\ \left| \frac{c\ m_i\ m_j}{\frac{z_i(\hat{\beta}_0)}{\sigma}\ \frac{z_j(\hat{\beta}_0)}{\sigma}} \right|\ \right],$$

where $\hat{\beta}_0$ is an initial estimate of β_0, the true parameter. m_i is defined as

$$m_i = \psi\left[\frac{b}{(\underline{x}_i - \underline{\mu}_i)'\ \underline{S}^{-1}\ (\underline{x}_i - \underline{\mu}_i)} \right].$$

The denominator in m_i is a measure of the leverage of the i^{th} observation. $\underline{u}_i$ and $\underline{S}$ are location and covariance matrices of $\underline{X}$, respectively. The tuning constants b and c are cutoff points for outliers. The parameter σ^2, which is the variance of ϵ_i, rescales the residuals in the ψ function. The variance σ^2 can be estimated by MAD which is defined as:

$$\text{MAD} = 1.483\ \text{med}\ |\ y_i - \underline{x}_i'\ \hat{\beta}_0 - \text{med}(\ y_i - \underline{x}_i'\ \hat{\beta}_0\ )\ |\ .$$

The MAD is a consistent estimate of σ^2 for normal errors, see Rousseeuw and Leroy(1987). In Chapter VI, we will discuss these constants in details.

From now, we assume that we have an initial estimate, and hence a set of weights of the above form. Note that the weights are always between 0 and 1. In order to discuss the geometry of our estimation, we first define the function:

$$\|\underline{u}\|_R = \sum_{i<j} b_{ij}\ |\ u_j - u_i\ |\ .$$

The function $\|\underline{u}\|_R$ is a pseudo norm which has the properties listed below.

Theorem 2.1

For all vectors $\underline{u}$, $\underline{v}$ and scaler k, we have the following properties:

1. $\| \underline{u} + \underline{v} \|_R \leq \| \underline{u} \|_R + \| \underline{v} \|_R$.

2. $\| k \underline{u} \|_R = |k| \, \| \underline{u} \|_R$.

3. $\| \underline{u} \|_R \geq 0$.

4. $\| \underline{u} \|_R = 0$, then $u_i = u_j$, for all i and j.

5. $\| \underline{u} + k \underline{1} \|_R = \| \underline{u} \|_R$, for all $k \in \mathbb{R}$, $\underline{1}$ is the vector whose components are all ones.

The proof of these properties is straight forward. In terms of this norm, the distance between two vectors $\underline{u}$ and $\underline{v}$ is $\| \underline{u} - \underline{v} \|_R$. Our estimate of $\underline{\beta}$ is $\hat{\underline{\beta}}$ which minimizes $\| \underline{Y} - \underline{X} \, \underline{\beta} \|_R$, the distance between $\underline{Y}$ and column space of $\underline{X}$. The existence of $\hat{\underline{\beta}}$ will be proved in the Theorem 2.3 below.

It is easy to see that the dispersion function $D(\underline{\beta})$ is the pseudo norm defined above. In the next theorem, we will use this pseudo norm and its associated properties to prove that this dispersion function $D(\underline{\beta})$ is a continuous and convex function of $\underline{\beta}$.

Theorem 2.2

$D(\underline{\beta})$ is a continuous and convex function of $\underline{\beta}$.

Proof.

The convexity of $D(\underline{\beta})$ follows from the following inequality:

$$D[\, t \underline{\beta}_1 + (1 - t) \underline{\beta}_2] = \| \underline{Y} - \underline{X} [\, t \underline{\beta}_1 + (1 - t) \underline{\beta}_2] \|_R =$$

$$= \| t \underline{Y} + (1 - t) \underline{Y} - t \underline{X} \underline{\beta}_1 - (1 - t) \underline{X} \underline{\beta}_2] \|_R \leq$$

$$\leq t \| \underline{Y} - \underline{X} \underline{\beta}_1 \|_R + (1 - t) \| \underline{Y} - \underline{X} \underline{\beta}_2 \|_R \leq$$

$\leq t D(\underline{\beta}_1) + (1 - t) D(\underline{\beta}_2)$,

for all $t \in [0, 1]$.

So $D[t \underline{\beta}_1 + (1 - t) \underline{\beta}_2] \leq t D(\underline{\beta}_1) + (1 - t) D(\underline{\beta}_2)$, for all $t \in [0, 1]$.

By the definition, $D(\underline{\beta})$ is a convex function of $\underline{\beta}$.

Because $D(\underline{\beta})$ is sum of absolute value functions, $D(\underline{\beta})$ is continuous. □

The existence of a minimizing value of the dispersion function $D(\underline{\beta})$ is a consequence of the above theorem. We will state it in the following theorem.

Theorem 2.3

A minimizing value of $D(\underline{\beta})$ always exists.

Proof.

Since $D(\underline{\beta})$ is continuous and convex, a minimizing value of $D(\underline{\beta})$ always exists. □

2.3 The Gradient of $D(\underline{\beta})$

Denote the negative gradient of the dispersion function $D(\underline{\beta})$ by

$$S(\underline{\beta}) = -\nabla D(\underline{\beta}).$$

It is more convenient to use the gradient $S(\underline{\beta})$ than the dispersion function $D(\underline{\beta})$ in theoretical discussion. Based on the definition above, we can easily prove the next theorem.

Theorem 2.4

$$S(\underline{\beta}) = \sum_{i<j} b_{ij} \operatorname{sgn}(z_j - z_i)(\underline{x}_j - \underline{x}_i).$$

Note that the above gradient can also be expressed as:

$$S(\underline{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\, \text{sgn}(\, z_j - z_i\,)\,(\, \underline{x}_j - \bar{\underline{x}}\,)\,.$$

In terms of the gradient, $\hat{\underline{\beta}}$ is a solution of $\underline{S}(\underline{\beta}) \doteq \underline{0}$. These equations are the analogues of the normal equations which define the least squares estimates.

Corollary 2.5

$\hat{\underline{\beta}}$ is a solution of $\underline{S}(\underline{\beta}) \doteq \underline{0}$.

Assumption A1

$\hat{\underline{\beta}}$ is regression and scale equivalent.

Under Assumption A1, $\hat{\underline{\beta}}$ is also regression and scale equivalent as the following theorem shows.

Theorem 2.6

Let $z_i(\underline{\beta}) = y_i - \underline{x}_i'\,\underline{\beta}$.

Then

1. $\hat{\underline{\beta}}(\underline{Z}(\underline{\beta} + \underline{\theta})) = \hat{\underline{\beta}}(\underline{Z}(\underline{\beta})) + \underline{\theta}$, for all $\underline{\theta} \in \mathbb{R}^p$.

2. $\hat{\underline{\beta}}(c\underline{Y}) = c\,\hat{\underline{\beta}}(\underline{Y})$, for all $c > 0$.

Proof.

$$S(\underline{\beta} + \underline{\theta}) = \sum_{i<j} b_{ij}\, \text{sgn}\{[(\, y_j - \underline{x}_j'\,\underline{\theta}\,) - \underline{x}_j'\,\underline{\beta}\,] - [(\, y_i - \underline{x}_i'\,\underline{\theta}\,) - \underline{x}_i'\,\underline{\beta}\,]\}\,(\, \underline{x}_j - \underline{x}_i\,)$$

$= 0$. □

Remark 2.7

The regression equivalence implies that in theoretical discussions, we can assume that

$\underline{\beta} = \underline{0}$ without loss of generality.

CHAPTER III

ASYMPTOTIC LINEARITY

In 1983, Sievers proposed a linear model with weights b_{ij} being functions of $\underline{X}$. As we discussed early, Naranjo and Hettmansperger(1990) proved that for appropriate weights, the influence function of the Sievers's general rank estimates is bounded in both the X and Y spaces, and the breakdown point is less than or equal to ⅓.

Because such weights b_{ij} downweight outliers indiscriminately in X space and some leverage points could be good points, we will use the weights which are functions of both $\underline{X}$ and $\underline{Y}$ in this thesis. With these weights, we will be able to only downweight bad leverage points and keep good leverage points. The weight functions also provide us the possibility of obtaining robust estimates with high efficiency.

In this chapter, we will discuss the asymptotic linearity with b_{ij} being function of both $\underline{X}$ and $\underline{Y}$. For convenience, we assume that the true parameter $\underline{\beta}_0 = \underline{0}$ without loss of generality. It will also be convenient to reparameterize the model as $\underline{\beta} = \dfrac{\underline{\Delta}}{\sqrt{n}}$ in this chapter.

As usual, before deriving the asymptotic linearity we first define several notations which will be used through rest part of this thesis. These notation will be proven useful for our theoretical development later.

Definition 3.1

1. $D_c = \{ \underline{\Delta} : |\Delta_k| \le c,\ k = 1, \ldots, p \}$, where $c > 0$.

2. $W_{ij} = \dfrac{\operatorname{sgn}(z_j - z_i) - \operatorname{sgn}(y_j - y_i)}{2}$, where $z_j = y_j - \underline{x}_j' \dfrac{\underline{\Delta}}{\sqrt{n}}$.

3. $\underline{R}(\underline{\Delta})^{p\times 1} = n^{-\frac{3}{2}} \left[\sum_{i<j} b_{ij} (\underline{x}_j - \underline{x}_i) W_{ij} + \underline{C}_n \dfrac{\underline{\Delta}}{\sqrt{n}} \right]$.

4. $\underline{C}_n^{p\times p} = [c_{kl}]$, $c_{kl} = \sum_{i<j} \gamma_{ij} b_{ij} (x_{jk} - x_{ik}) (x_{jl} - x_{il})$,

where $\gamma_{ij} = \dfrac{B_{ij}'(0)}{E_{\underline{\beta}_0}(b_{ij})}$, $B_{ij}(t) = E_{\underline{\beta}_0}[b_{ij} I(0 < y_i - y_j < t)]$. The existence of

$B_{ij}'(0)$ will be proven below.

5. $\underline{X}_c = (I_n - \frac{1}{n} J_n) \underline{X}$, where J_n is an $n\times n$ matrix of ones.

6. $\|u\| = \sqrt{\sum_{i=1}^{p} | u_i |}$, for all $\underline{u} \in \mathbb{R}^p$, for all $p \in \mathbb{N}$.

Before deriving the asymptotic linearity, we first need to show several lemmas. These lemmas will be used in the development of the asymptotic linearity later this chapter. The next lemma can be found in most calculus text books.

Lemma 3.2

Suppose that functions $g(y, t)$ and $g_t'(y, t)$ are continuous on $[a, b] \times [c, d]$, and for all $t \in [c, d]$, we have $y = \alpha_i(t) \in [a, b]$ and $\alpha_i'(t)$ exists, $i = 1, 2$.

Then we have the following:

$\dfrac{d}{dt} \int_{\alpha_1(t)}^{\alpha_2(t)} g(y, t)\, dy =$

$$= \int_{\alpha_1(t)}^{\alpha_2(t)} g_t'(y, t)\, dy + \alpha'_2(t) \cdot g[\alpha_2(t), t] - \alpha'_1(t) \cdot g[\alpha_1(t), t].$$

Further assumptions on the weights and the density function of the random variables are needed, and they will be stated below.

Assumption A2

1. b_{ij} are continuous with respect to y_i and y_j, for all i and j.
2. f is continuous and bounded.

Recall that in Definition 3.1, we defined:

$$B_{ij}(t) = E_{\beta_0}[\, b_{ij}\, I(0 < y_i - y_j < t)\,].$$

The derivative of $B_{ij}(t)$ is continuous under the above assumption. In the next two theorems, we will discuss some properties of $B_{ij}(t)$. These properties will be proven very useful in our theoretical development later.

Theorem 3.3

Assume A2 holds. Then $B_{ij}'(t)$ is continuous, for all t.

Proof.

By A2 and Lemma 3.2. □

Theorem 3.4

Assume A2 holds, and let $t_{ij} = (\underline{x}_j - \underline{x}_i)' \frac{\underline{\Delta}}{\sqrt{n}}$

Then

There exists constants ξ_{ij} such that $|\xi_{ij}| < |t_{ij}|$ and $E_{\beta_0}(b_{ij} W_{ij}) = -t_{ij} B_{ij}'(\xi_{ij})$.

Proof.

Because $W_{ij} = 1, -1, 0$ as $t_{ij} < y_j - y_i < 0$, $0 < y_j - y_i < t_{ij}$, otherwise, we have the following equation:

$$E_{\beta_0}(\,b_{ij}\,W_{ij}\,) = \int_{t_{ij} < y_j - y_i < 0} b_{ij}\, f_{\underline{Y}}\, d\underline{Y} - \int_{0 < y_j - y_i < t_{ij}} b_{ij}\, f_{\underline{Y}}\, d\underline{Y}\,.$$

When $t_{ij} > 0$, $E_{\beta_0}(\,b_{ij}\,W_{ij}\,) = -\,B_{ij}(t_{ij}) = B_{ij}(0) - B_{ij}(t_{ij}) = -\,t_{ij}\,B_{ij}'(\xi_{ij})\,.$

The result is the same, when $t_{ij} < 0$. □

Followings are assumptions for the design matrix and for the regression constants. Recall that in the definition 3.1, we have defined the centered design matrix:

$$\underline{X}_c = (\,\underline{I}_n - \frac{1}{n}\,\underline{J}_n\,)\,\underline{X}\,,$$

where J_n is an $n{\times}n$ matrix of ones.

<u>Assumption A3</u>

There exists a positive definite matrix Γ such that $\frac{1}{n}\,\underline{X}'_c\,\underline{X}_c \to \Gamma$.

<u>Assumption A4</u>

Noether's condition holds, i.e. $\dfrac{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}{\max_{1 \le i \le n} (x_{ik} - \bar{x}_k)^2} \to \infty$, as $n \to \infty$, for all k.

<u>Remark 3.5</u>

Note that Assumption A3 implies that $\frac{1}{n} \sum_{i=1}^{n} (\,x_{ik} - \bar{x}_k\,)^2 \to M < \infty$, for some $M > 0$, for all k. Therefore, by A3 - A4, we have $\frac{1}{n} \max_{1 \le i \le n} (\,x_{ik} - \bar{x}_k\,)^2 \to 0$, as $n \to \infty$, for all k. Hence, we can conclude that $\frac{1}{\sqrt{n}} \max_{1 \le i \le n} |\,x_{ik} - \bar{x}_k\,| \to 0$, as $n \to \infty$, for all k.

We also need to assume that the initial estimate of $\underset{\sim}{\beta}$ is $\sqrt{n}$ - consistent, and it will be stated in the next assumption.

Assumption A5

$$\sqrt{n}\,(\hat{\underset{\sim}{\beta}}_0 - \underset{\sim}{\beta}_0) \xrightarrow{d} N(\underset{\sim}{0}, \Xi),$$ where $\underset{\sim}{\beta}_0$ is the true parameter and Ξ is a positive definite matrix.

Since the weights are functions of regression constants, the response variables, and the initial estimate of $\underset{\sim}{\beta}$, it will be convenient to use the notation below in our theoretical proving:

$$b_{ij} = g(\underset{\sim}{x}_i, \underset{\sim}{x}_j, y_i, y_j, \hat{\underset{\sim}{\beta}}_0) = g_{ij}(\hat{\underset{\sim}{\beta}}_0).$$

Using the above notation, we have the following lemma.

Lemma 3.6

$$g_{ij}(\hat{\underset{\sim}{\beta}}_0) = g_{ij}(\underset{\sim}{\beta}_0) + \nabla g_{ij}'(\underset{\sim}{\xi}) \cdot (\hat{\underset{\sim}{\beta}}_0 - \underset{\sim}{\beta}_0),$$ $\underset{\sim}{\xi}$ is in between $\hat{\underset{\sim}{\beta}}_0$ and $\underset{\sim}{\beta}_0$.

Proof.

By Mean-Value Theorem (P.355 of Apostol, 1974). □

Assumption A6

$$\frac{1}{n}\sum_{i=1}^{n} |x_{ik} - \bar{x}_k| = O(1), \quad \text{for all } k.$$

In the next three lemmas, Lemma 3.7 through Lemma 3.9, we will discuss several expected values which will be used in our development of the asymptotic linearity in Theorem 3.10 later.

Lemma 3.7

Assume A2 - A4 hold, and let $t_{ij} = (\underline{x}_j - \underline{x}_i)' \frac{\Delta}{\sqrt{n}}$.

Then

1. For all $\varepsilon > 0$, there exists an N_1 such that $| E(g_{ij}\, g_{li}\, W_{ij}\, W_{li}) | < \varepsilon$, for all $n > N_1$.

2. For all $\varepsilon > 0$, there exists an N_2 such that $| E(g_{ij}\, W_{ij})| < \varepsilon$, for all $n > N_2$.

Proof.

Without loss of generality, we assume that: $t_{ij} > 0$ and $t_{li} > 0$, where indices i , j , l are all different. Then we have:

$| E(g_{ij}\, g_{li}\, W_{ij}\, W_{li}) | =$

$= | E[g_{ij}\, g_{li}\; I(0 < y_j - y_i < t_{ij})\; I(0 < y_i - y_l < t_{li})] | \leq$

$\leq | E[I(0 < y_j - y_i < t_{ij})\; I(0 < y_i - y_l < t_{li})] | =$

$$= \int_{-\infty}^{+\infty} \int_{y_i - t_{li}}^{y_i} \int_{y_i}^{y_i + t_{ij}} f_i\, f_j\, f_l\; dy_j\, dy_l\, dy_i .$$

Therefore by A2 - A4, for all $\varepsilon > 0$, there exists an N_1 such that for all $n > N_1$,

$| E(g_{ij}\, g_{li}\, W_{ij}\, W_{li}) | < \varepsilon$, Hence the first condition is true.

For the second condition, we can again assume without loss of generality that $t_{ij} > 0$.

Then we have the following inequality:

$| E(g_{ij}\, W_{ij}) | = | E[g_{ij}\; I(0 < y_j - y_i < t_{ij})] | \leq$

$\leq | E[I(0 < y_j - y_i < t_{ij})] | =$

$$= \int_{-\infty}^{+\infty} \int_{y_i}^{y_i + t_{ij}} f_i\, f_j\; dy_j\, dy_i .$$

So once again by A2 - A4, for all $\varepsilon > 0$, there exists an N_2 such that

$| E(g_{ij} W_{ij}) | < \varepsilon$, for all $n > N_2$. Hence the second condition is true. □

Lemma 3.8

Assume A2 - A5 hold. Let $t_{ij} = (\underline{x}_j - \underline{x}_i)' \dfrac{\underline{\Delta}}{\sqrt{n}}$, and

$u_{ij} = \nabla g_{ij}(\underline{\xi})' \cdot \sqrt{n} (\hat{\underline{\beta}}_0 - \underline{\beta}_0)$.

Then

1. For all $\varepsilon > 0$, there exists an N_1 such that for all $n > N_1$,

$$\left| n^{-\frac{7}{2}} \sum_{\substack{i=1 \\ (i,j) \neq (l,m)}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) E(u_{ij} g_{lm} W_{ij} W_{lm}) \right| < \varepsilon .$$

2. For all $\varepsilon > 0$, there exists an N_2 such that for all $n > N_2$,

$$\left| n^{-\frac{7}{2}} \sum_{\substack{i=1 \\ (i,j) \neq (l,m)}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) E(u_{ij} W_{ij}) E(g_{lm} W_{lm}) \right| < \varepsilon .$$

Proof.

1. Without loss of generality, we assume that $t_{ij} > 0$ and $t_{lm} > 0$, where indices i , j , l , m are all different. Define $B_{ij,lm}(t_{ij} , t_{lm}) = E(u_{ij} g_{lm} W_{ij} W_{lm})$, then

$E(u_{ij} g_{lm} W_{ij} W_{lm}) = E[u_{ij} g_{lm} I(0 < y_j - y_i < t_{ij}) I(0 < y_m - y_l < t_{lm})] =$

$$= E \int_{-\infty}^{+\infty} \int_{y_l}^{y_l + t_{lm}} \int_{-\infty}^{+\infty} \int_{y_i}^{y_i + t_{ij}} u_{ij} g_{lm} f_i f_j f_l f_m \, dy_j \, dy_i \, dy_m \, dy_l = B_{ij,lm}(t_{ij} , t_{lm}) =$$

$= [t_{ij} , t_{lm}]' \cdot \nabla B_{ij,lm}(\xi_{ij} , \xi_{lm})$, $|\xi_{ij}| \le |t_{ij}|$ and $|\xi_{lm}| \le |t_{lm}|$.

The last statement holds, since $\nabla B_{ij,lm}(\xi_{ij} , \xi_{lm})$ is continuous with $\nabla B_{ij,lm}(0 , 0) = \underline{0}$. Therefore by A3 - A4, we obtain the conclusion.

2. Without loss of generality, we assume that $t_{ij} > 0$, where indices i, j are different.

   Define $BU_{ij}(t_{ij}) = E(u_{ij} W_{ij})$, then

   $E(u_{ij} W_{ij}) = E[u_{ij} I(0 < y_j - y_i < t_{ij})] =$

   $$= E \int_{-\infty}^{+\infty} \int_{y_i}^{y_i + t_{ij}} u_{ij} f_i f_j \, dy_j \, dy_i =$$

   $= BU_{ij}(t_{ij}) = t_{ij} BU_{ij}'(\xi_{ij})$, where $|\xi_{ij}| \le |t_{ij}|$.

   The above statement holds because $u_{ij} = \nabla g_{ij}'(\underline{\xi}) \cdot \sqrt{n} (\hat{\underline{\beta}}_0 - \underline{\beta}_0)$. Therefore

   $BU_{ij}'(\xi_{ij})$ is bounded by A5.

   Similarly without loss of generality, we assume that $t_{lm} > 0$, where indices l and m are different. Define $BG_{lm}(t_{lm}) = E(g_{lm} W_{lm})$, then

   $E(g_{lm} W_{lm}) = E[g_{lm} I(0 < y_m - y_l < t_{lm})] =$

   $$= E \int_{-\infty}^{+\infty} \int_{y_l}^{y_l + t_{lm}} g_{lm} f_l f_m \, dy_m \, dy_l =$$

   $= BG_{lm}(t_{lm}) = t_{lm} BG_{lm}'(\xi_{lm})$, where $|\xi_{lm}| \le |t_{lm}|$.

   Note that we used $0 \le g_{lm} \le 1$. Therefore, $BG'_{lm}(\xi_{lm})$ is bounded.

   Hence the result follows from Assumption A3-A4. □

Lemma 3.9

Assume A2 - A5. Let $t_{ij} = (\underline{x}_j - \underline{x}_i)' \frac{\Delta}{\sqrt{n}}$, and

$u_{ij} = \nabla g_{ij}'(\underline{\xi}) \cdot \sqrt{n} (\hat{\underline{\beta}}_0 - \underline{\beta}_0)$.

Then

1. For all $\varepsilon > 0$, there exists an N_1 such that $| E(u_{ij} u_{lm} W_{ij} W_{lm}) | < \varepsilon$,

for all $n > N_1$.

2. For all $\varepsilon > 0$, there exists an N_2 such that $| E(u_{ij} W_{ij}) | < \varepsilon$, for all $n > N_2$.

Proof.

1. Without loss of generality, we assume that $t_{ij} > 0$ and $t_{lm} > 0$, where indices i, j, l, m are all different.

   Then $| E(u_{ij} u_{lm} W_{ij} W_{lm}) | =$

   $= | E[u_{ij} u_{lm} I(0 < y_j - y_i < t_{ij}) I(0 < y_m - y_l < t_{lm})] | =$

   $$= | E \int_{-\infty}^{+\infty} \int_{y_i}^{y_i+t_{ij}} \int_{-\infty}^{+\infty} \int_{y_l}^{y_l+t_{lm}} u_{ij} u_{lm} f_i f_j f_l f_m \, dy_m \, dy_l \, dy_j \, dy_i | .$$

   So the result holds by A5.

2. Without loss of generality, we assume that $t_{ij} > 0$, where indices i and j are different.

   Then $| E(u_{ij} W_{ij}) | = | E[u_{ij} I(0 < y_j - y_i < t_{ij})] | =$

   $$= | E \int_{-\infty}^{+\infty} \int_{y_i}^{y_i+t_{ij}} u_{ij} f_i f_j \, dy_j \, dy_i | .$$

   By A5, the result holds. □

Now, we are ready to derive our main result of this chapter, the asymptotic linearity, in the next theorem. Recall the definition of $R(\underline{\Delta})$ in the beginning of this chapter, we can rewrite the process $R(\underline{\Delta})$ as:

$$\underline{R}(\underline{\Delta})^{p\times 1} = n^{-\frac{3}{2}} \sum_{i<j} b_{ij} (\underline{x}_j - \underline{x}_i) W_{ij} + n^{-\frac{3}{2}} \sum_{i<j} \underline{C}_n \frac{\Delta}{\sqrt{n}} .$$

It is easy to see that the first term is a function of $S(\underline{\beta})$, and the second term is a linear function of $\frac{\Delta}{\sqrt{n}}$. If we can show that $\| R(\underline{\Delta}) \| \xrightarrow{P} 0$ as $n \to \infty$, then the first term is approximately equal to the second term, the linear function of $\frac{\Delta}{\sqrt{n}}$, when n is sufficiently large.

Theorem 3.10

Assume A2-A6 hold. Let $\underline{\beta} = \frac{\Delta}{\sqrt{n}}$.

Then

$$\sup_{\Delta \in D_c} \| \underline{R}(\underline{\Delta}) \| \xrightarrow{P} \underline{0}.$$

Proof.

Let $t_{ij}(\underline{\Delta}) = (\underline{x}_j - \underline{x}_i)' \frac{\Delta}{\sqrt{n}}$, then for all k,

$$R_k(\underline{\Delta}) = n^{-\frac{3}{2}} [\sum_{i<j} b_{ij} (x_{jk} - x_{ik}) W_{ij} + \sum_{i<j} \gamma_{ij} b_{ij} (x_{jk} - x_{ik}) t_{ij}] =$$

$$= n^{-\frac{3}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)(b_{ij} W_{ij} + \gamma_{ij} t_{ij} b_i).$$

First, we will show that for all $\varepsilon > 0$, there exists an N such that for all $n > N$, $E_{\underline{\beta}_0}(R_k) < \varepsilon$, for all $\underline{\Delta} \in D_c$ and for all k.

For all k, $E_{\underline{\beta}_0}(R_k) = n^{-\frac{3}{2}} \sum_{i<j} (x_{jk} - x_{ik}) [E_{\underline{\beta}_0}(b_{ij} W_{ij}) + \gamma_{ij} t_{ij} E_{\underline{\beta}_0}(b_{ij})] =$

$$= n^{-\frac{3}{2}} \sum_{i<j} (x_{jk} - x_{ik}) t_{ij} [B_{ij}'(0) - B_{ij}'(\xi_{ij})] \le$$

$$\leq n^{-\frac{3}{2}} \sqrt{\sum_{i<j} (x_{jk} - x_{ik})^2} \sqrt{\sum_{i<j} t_{ij}^2} \sup_{i<j} | B_{ij}'(0) - B_{ij}'(\xi_{ij}) | ,$$ by Cauchy inequality.

Because $\sum_{i<j} (x_{jk} - x_{ik})^2 = n \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2 , \quad \sum_{i<j} t_{ij}^2 = \underline{\Delta}' \underline{X}'_c \underline{X}_c \underline{\Delta}$ (3.1)

and for all $\varepsilon^* > 0 , \quad \sup_{i<j} | B_{ij}'(0) - B_{ij}'(\xi_{ij}) | < \varepsilon^*$,

for all $\underline{\Delta} \in D_c$, for n sufficiently large.

Therefore for all $\varepsilon > 0$, there exists an N such that for all $n > N$,

$$|E_{\beta_0}(R_k)| \leq n^{-\frac{3}{2}} \sqrt{n \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} \sqrt{\sum_{i=1}^{n} t_{ij}^2} \sup_{i<j} | B_{ij}'(0) - B_{ij}'(\xi_{ij}) | < \varepsilon .$$

So for all $\varepsilon > 0$, there exists an N such that for all $n > N$, $E(R_k) < \varepsilon$, for all $\underline{\Delta} \in D_c$, by A2 - A3 and Theorem 3.3.

Next, we will show that for all $\varepsilon > 0$, there exists an N such that for all $n > N$, $Var_{\beta_0}(R_k) < \varepsilon$, for all $\underline{\Delta} \in D_c$, for all k.

For all k, $Var_{\beta_0}(R_k) = n^{-3} \; Var_{\beta_0}[\sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k) (b_{ij} W_{ij} + \gamma_{ij} t_{ij} b_{ij})] =$

$$= n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \; Var_{\beta_0}(b_{ij} W_{ij} + \gamma_{ij} t_{ij} b_{ij}) +$$

$$+ n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{\substack{m=1 \\ (i,j) \neq (l,m)}}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) \cdot$$

$$\cdot \; Cov(b_{ij} W_{ij} + \gamma_{ij} t_{ij} b_{ij} , b_{lm} W_{lm} + \gamma_{lm} t_{lm} b_{lm}) =$$

$$= n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 [Var(b_{ij} W_{ij}) + \gamma_{ij}^2 t_{ij}^2 Var(b_{ij}) +$$

$$+ 2 \gamma_{ij} t_{ij} Cov(b_{ij} W_{ij} , b_{ij})] +$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{\substack{m=1 \\ (i,j)\neq(l,m)}}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)[\,\mathrm{Cov}(b_{ij}W_{ij},\ b_{lm}W_{lm}) +$$

$$+ \gamma_{lm} t_{lm} \mathrm{Cov}(b_{ij}W_{ij},\ b_{lm}) + \gamma_{ij} t_{ij} \mathrm{Cov}(b_{ij},\ b_{lm}W_{lm}) +$$

$$+ \gamma_{ij}\gamma_{lm} t_{ij} t_{lm} \mathrm{Cov}(b_{ij}, b_{lm})\,].$$

For the first part, we have:

Because $\mathrm{Var}(b_{ij}W_{ij})$, $\mathrm{Var}(b_{ij})$, and $\mathrm{Cov}(b_{ij}W_{ij}, b_{ij})$ are bounded, therefore for all $\varepsilon^* > 0$, there exists an N such that for all $n > N$,

$$1^{st}\text{ term} = n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \mathrm{Var}(b_{ij}W_{ij}) < \varepsilon^*,$$

$$2^{nd}\text{ term} = n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \gamma_{ij}^2 t_{ij}^2 \mathrm{Var}(b_{ij}) \le$$

$$\le n^{-2} \max_{i,j}[\,\gamma_{ij}^2 t_{ij}^2 \mathrm{Var}(b_{ij})\,] \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 < \varepsilon^*,$$

$$3^{rd}\text{ term} = n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2\, 2\gamma_{ij} t_{ij} \mathrm{Cov}(b_{ij}W_{ij},\ b_{ij})\,] \le$$

$$\le 2n^{-2} \max_{i,j} |\,\gamma_{ij} t_{ij} \mathrm{Cov}(b_{ij}W_{ij},\ b_{ij})\,| \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 < \varepsilon^*.$$

For the second part, we will use Lemma 3.6 in the following proof:

For the term $n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{\substack{m=1 \\ (i,j)\neq(l,m)}}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k) \mathrm{Cov}(b_{ij}W_{ij},\ b_{lm}W_{lm}) =$

$$= n^{-3} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{\substack{m=1 \\ (i,j)\neq(l,m)}}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\{\,\mathrm{Cov}[\,g_{ij}(\underline{\beta}_0)W_{ij},\ g_{lm}(\underline{\beta}_0)W_{lm}\,] +$$

$$+ \mathrm{Cov}[\,\nabla g_{ij}'(\underline{\xi}) \cdot (\hat{\underline{\beta}}_0 - \underline{\beta}_0)W_{ij},\ g_{lm}(\underline{\beta}_0)W_{lm}\,] +$$

$$+ \mathrm{Cov}[\,g_{ij}(\underline{\beta}_0)W_{ij},\ \nabla g_{lm}'(\underline{\xi}) \cdot (\hat{\underline{\beta}}_0 - \underline{\beta}_0)W_{lm}\,] +$$

$+ \operatorname{Cov}[\nabla g_{ij}'(\xi) \cdot (\hat{\beta}_0 - \beta_0) W_{ij} , \nabla g_{lm}'(\xi) \cdot (\hat{\beta}_0 - \beta_0) W_{lm}] \}.$

$$1^{st} \text{ term } = n^{-3} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) \operatorname{Cov}[g_{ij}(\beta_0) W_{ij} , g_{lm}(\beta_0) W_{lm}] .$$

$\operatorname{Cov}[g_{ij}(\beta_0) W_{ij} , g_{lm}(\beta_0) W_{lm}] = 0$, when indices i, j, l, m are all different.

If at least two indices are the same, for example, $i = m$, then

$$1^{st} \text{ term } = n^{-3} \sum_{i=1}^{n} \sum_{\substack{j=1\\j\neq l}}^{n} \sum_{l=1}^{n} (x_{jk} - \bar{x}_k) (x_{ik} - \bar{x}_k) \operatorname{Cov}[g_{ij}(\beta_0) W_{ij} , g_{li}(\beta_0) W_{li}] \leq$$

$$\leq \max_{i,j,l} | \operatorname{Cov}[g_{ij}(\beta_0) W_{ij} , g_{li}(\beta_0) W_{li}] | \; (n^{-1} \sum_{i=1}^{n} | x_{jk} - \bar{x}_k |)^2 .$$

Because $\operatorname{Cov}[g_{ij}(\beta_0) W_{ij} , g_{li}(\beta_0) W_{li}] =$

$= E(g_{ij} g_{li} W_{ij} W_{li}) - E(g_{ij} W_{ij}) E(g_{li} W_{li})$,

according to Lemma 3.7, A2 - A4 and A6, for all $\varepsilon^* > 0$, there exists an N^* such that

1^{st} term $< \varepsilon^*$ for all $n > N^*$.

Similarly, it can be shown that 1^{st} term $< \varepsilon^*$ in other cases.

By A5, letting $\frac{u_{ij}}{\sqrt{n}} = \nabla g_{ij}'(\xi) \cdot (\hat{\beta}_0 - \beta_0)$, then

$$2^{nd} \text{ term } = n^{-\frac{7}{2}} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) \operatorname{Cov}[u_{ij} W_{ij} , g_{lm}(\beta_0) W_{lm}] =$$

$$= n^{-\frac{7}{2}} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} (x_{jk} - \bar{x}_k) (x_{mk} - \bar{x}_k) [E(u_{ij} g_{lm} W_{ij} W_{lm}) -$$

$$- E(u_{ij} W_{ij}) E(g_{lm} W_{lm})] .$$

By Lemma 3.8, for all $\varepsilon^* > 0$, there exists an N^* such that 2^{nd} term $< \varepsilon^*$ for all $n > N^*$.

Similarly, we can show that 3^{rd} term $< \varepsilon^*$.

$$4^{th}\text{ term} = n^{-3} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\,\mathrm{Cov}\!\left(\frac{u_{ij}}{\sqrt{n}} W_{ij},\ \frac{u_{lm}}{\sqrt{n}} W_{lm}\right) =$$

$$= n^{-4} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\,[\,E(u_{ij} u_{lm} W_{ij} W_{lm}) -$$

$$- E(u_{ij} W_{ij})\, E(u_{lm} W_{lm})\,].$$

By Lemma 3.9, for all $\varepsilon^* > 0$, there exists an N^* such that 4^{th} term $< \varepsilon^*$ for all $n > N^*$.

Hence by the proof above, for all $\varepsilon^{**} > 0$, there exists an N^{**} such that for all $n > N^{**}$, the term

$$n^{-3} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\,\mathrm{Cov}(b_{ij} W_{ij},\ b_{lm} W_{lm}) < \varepsilon^{**}.$$

Similar to the proof above, we can also show that:

For all $\varepsilon^* > 0$, there exists an N^* such that for all $n > N^*$, the term

$$n^{-3} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\,\gamma_{lm} t_{lm}\,\mathrm{Cov}(b_{ij} W_{ij},\ b_{lm}) =$$

$$= n^{-3} \sum_{\substack{i=1\\(i,j)\neq(l,m)}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\,\gamma_{lm} t_{lm}\,[\,\mathrm{Cov}(g_{ij} W_{ij},\ g_{lm}) +$$

$$+ \frac{1}{\sqrt{n}}\mathrm{Cov}(g_{ij} W_{ij},\ u_{lm}) + \frac{1}{\sqrt{n}}\mathrm{Cov}(u_{ij} W_{ij},\ g_{lm}) + \frac{1}{n}\mathrm{Cov}(u_{ij} W_{ij},\ u_{lm}) <$$

$$< \varepsilon^*.$$

Next, we look at the term:

$$n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n}{}_{(i,j)\neq(l,m)} (x_{jk} - \bar{x}_k)(x_{mk} - \bar{x}_k)\, \gamma_{ij}\, \gamma_{lm}\, t_{ij}\, t_{lm}\, \mathrm{Cov}(b_{ij}, b_{lm}) .$$

Because $\mathrm{Cov}(b_{ij}, b_{lm}) =$

$$= \mathrm{Cov}(g_{ij}, g_{lm}) + \frac{1}{\sqrt{n}} \mathrm{Cov}(g_{ij}, u_{lm}) + \frac{1}{\sqrt{n}} \mathrm{Cov}(u_{ij}, g_{lm}) + \frac{1}{n} \mathrm{Cov}(u_{ij}, u_{lm}),$$

therefore by A3, for all $\varepsilon^* > 0$, there exists an N^* such that the above term $< \varepsilon^*$ for all $n > N^*$.

Hence for all $\varepsilon > 0$, there exists an N such that $\mathrm{Var}(R_k) < \varepsilon$, for all $n > N$, for all $\underline{\Delta} < D_c$, for all k.

By the proof above, $\| R_k \| \xrightarrow{p} 0$, for all $\underline{\Delta} \in D_c$, for all k.

Therefore $\sup_{\Delta \in D_c} \| \underline{R}(\underline{\Delta}) \| \leq p \sup_{\Delta \in D_c, k} \| R_k \| \xrightarrow{p} 0$. □

Remark 3.11

The Theorem 3.10 implies that $n^{-\frac{3}{2}} [\underline{S}(\underline{0}) - 2\, \underline{C}_n\, \underline{\beta}]$ is a linear approximation of $n^{-\frac{3}{2}} \underline{S}(\underline{\beta})$ for $\underline{\beta}$ in a local neighborhood of $\underline{0}$. Thus we have obtained an asymptotic linearity result.

The Remark 3.11 is the key result of this chapter. Since $\underline{S}(\underline{\beta})$ is a complicated function of $\underline{\beta}$, it is not easy to derive properties of $\hat{\underline{\beta}}$ based on $\underline{S}(\underline{\beta})$. On the other end, $[\underline{S}(\underline{0}) - 2\, \underline{C}_n\, \underline{\beta}]$ is a linear approximation of $n^{-\frac{3}{2}} \underline{S}(\underline{\beta})$ for $\underline{\beta}$ in a local neighborhood of $\underline{0}$. As we will see in the next chapter, we will use this asymptotic linearity to derive the asymptotic properties of $\hat{\beta}$.

CHAPTER IV

ASYMPTOTIC DISTRIBUTION OF $\hat{\underline{\beta}}$

In last chapter, we discussed the linear approximation of the gradient $\underline{S}(\underline{\beta})$. Based on this linear approximation, we will derive the asymptotic distribution of $\hat{\underline{\beta}}$, the estimates of the regression coefficients. The asymptotic distribution will play a very important role in the theoretical development of the testing and confidence regions for the estimates. If we assume the true parameter $\underline{\beta}_0 = \underline{0}$, and $\underline{\beta} = \frac{\underline{\Delta}}{\sqrt{n}}$, then we have the following approximation:

$$n^{-\frac{3}{2}} [\, \underline{S}(\underline{0}) - 2\, \underline{C}_n \frac{\underline{\Delta}}{\sqrt{n}} \,] \approx n^{-\frac{3}{2}} \underline{S}(\frac{\underline{\Delta}}{\sqrt{n}})$$

according to Theorem 3.10, when n is sufficiently large.

Because the process $\underline{S}(\frac{\underline{\Delta}}{\sqrt{n}})$ is approximately linear, it would seem then that the dispersion function $D(\frac{\underline{\Delta}}{\sqrt{n}})$ should be approximately a quadratic function. To see this, we let

$$\frac{\partial}{\partial \underline{\Delta}} Q(\frac{\underline{\Delta}}{\sqrt{n}}) = \frac{1}{\sqrt{n}} [\, 2\, \underline{C}_n \frac{\underline{\Delta}}{\sqrt{n}} - \underline{S}(\underline{0}) \,], \quad \frac{\partial}{\partial \underline{\Delta}} D(\frac{\underline{\Delta}}{\sqrt{n}}) = -\frac{1}{\sqrt{n}} \underline{S}(\frac{\underline{\Delta}}{\sqrt{n}}),$$

and define:

$$\frac{\partial Q^*(\underline{\Delta})}{\partial \underline{\Delta}} = \frac{1}{n}\frac{\partial}{\partial \underline{\Delta}} Q(\frac{\underline{\Delta}}{\sqrt{n}}), \qquad \frac{\partial D^*(\underline{\Delta})}{\partial \underline{\Delta}} = \frac{1}{n}\frac{\partial}{\partial \underline{\Delta}} D(\frac{\underline{\Delta}}{\sqrt{n}}).$$

Then we have a quadratic function:

$$Q(\frac{\underline{\Delta}}{\sqrt{n}}) = \frac{\underline{\Delta}'}{\sqrt{n}} \underline{C}_n \frac{\underline{\Delta}}{\sqrt{n}} - \sum_{i<j} b_{ij}\, \mathrm{sgn}(y_j - y_i)(\underline{x}_j - \underline{x}_i)' \frac{\underline{\Delta}}{\sqrt{n}} + D(\underline{0})$$

and

$$D^*(\underline{\Delta}) = \frac{1}{n} D(\frac{\underline{\Delta}}{\sqrt{n}}), \qquad Q^*(\underline{\Delta}) = \frac{1}{n} Q(\frac{\underline{\Delta}}{\sqrt{n}}).$$

Consequentially, we have the definition below.

Definition 4.1

$$Q^*(\underline{\Delta}) = n^{-2}\, \underline{\Delta}' \underline{C}_n \underline{\Delta} - n^{-\frac{3}{2}} \sum_{i<j} b_{ij}\, \mathrm{sgn}(y_j - y_i)(\underline{x}_j - \underline{x}_i)' \underline{\Delta} + n^{-1} D(\underline{0}).$$

We will state the following obvious conclusions as a lemma.

Lemma 4.2

1. $\hat{\underline{\Delta}}$ minimizes $D^*(\Delta)$ if and only if $\hat{\underline{\Delta}}$ minimizes $D(\frac{\underline{\Delta}}{\sqrt{n}})$.

2. $\tilde{\underline{\Delta}}$ minimizes $Q^*(\Delta)$ if and only if $\tilde{\underline{\Delta}}$ minimizes $Q(\frac{\underline{\Delta}}{\sqrt{n}})$.

Note that $\tilde{\underline{\Delta}}$ is not an estimate. It depends, as does Q, on the true parameter. However, it is a convenient tool for obtaining our results. The next theorem yields an asymptotic quadratic result for our dispersion function. It will eventually allow us to obtain the asymptotic distribution theory for $\hat{\underline{\Delta}}$. We will use the technique of Hettmansperger(1984) to prove the next theorem.

Theorem 4.3

Assume A2 - A6 hold.

Then

$P(\sup_{\Delta \in D_c} | Q^*(\underline{\Delta}) - D^*(\underline{\Delta}) | \geq \varepsilon) \to 0$, for all $\varepsilon > 0$, and for all $c > 0$.

Proof.

Assume the true parameter $\underline{\beta}_0 = \underline{0}$, and let $\underline{\beta} = \frac{\Delta}{\sqrt{n}}$.

Because $\frac{\partial Q^*}{\partial \Delta} - \frac{\partial D^*}{\partial \Delta} = 2\, n^{-\frac{3}{2}} [\sum_{i<j} b_{ij} (\underline{x}_j - \underline{x}_i) W_{ij} + \underline{C}_n \frac{\Delta}{\sqrt{n}}] = 2\, \underline{R}(\underline{\Delta})$,

by Theorem 3.10, for all $\varepsilon > 0$, we have:

$$P(\sup_{\Delta \in D_c} \| \frac{\partial Q^*(\Delta)}{\partial \Delta} - \frac{\partial D^*(\Delta)}{\partial \Delta} \| \geq \frac{\varepsilon}{pc}) \to 0 . \tag{4.1}$$

Let $\underline{\Delta}_t = t\, \underline{\Delta}$, for all $t \in [0, 1]$, then

$$\frac{d}{dt} [Q^*(\underline{\Delta}_t) - D^*(\underline{\Delta}_t)] = \sum_{k=1}^{p} \Delta_k (\frac{\partial Q^*}{\partial \Delta_{tk}} - \frac{\partial D^*}{\partial \Delta_{tk}}) .$$

According to (4.1) above, with high probability and for large n, we have:

$$\left| \frac{d}{dt} [Q^*(\Delta_t) - D^*(\Delta_t)] \right| \leq \| \Delta \| \sup_{\Delta \in D_c} \| \frac{\partial Q^*}{\partial \Delta} - \frac{\partial D^*}{\partial \Delta} \| < \| \Delta \| \frac{\varepsilon}{pc} \leq \varepsilon , \quad \text{a.e.}$$

Next, we define $h(t) = Q^*(\underline{\Delta}_t) - D^*(\underline{\Delta}_t)$, hence $h(t)$ is differentiable almost everywhere, and by the last result we have $| h'(t) | < \varepsilon$.

Hence $| h(1) - h(0) | \leq \int_0^1 | h'(t) | \, dt$.

Because $h(0) = 0$, and $| h'(t) | < \varepsilon$, for all t, $0 < t < 1$, we have $| h(1) | < \varepsilon$.

So $P(| Q^*(\underline{\Delta}) - D^*(\underline{\Delta}) | \geq \varepsilon) \to 0$, for all $\Delta \in D_c$.

So $P(\sup_{\Delta \in D_c} | Q^*(\underline{\Delta}) - D^*(\underline{\Delta}) | \geq \varepsilon) \to 0$. □

Because Q^* is a quadratic function of $\underline{\Delta}$, by differentiating, we are able to easily to obtain $\tilde{\underline{\Delta}}$ which is a minimizing value of Q^*. For future reference, we will state the result in the following corollary.

Corollary 4.4

Assume A2 - A6 hold, then

$$\tilde{\underline{\Delta}} = \frac{1}{2} \sqrt{n}\, \underline{C}_n^{-1} \sum_{i<j} b_{ij} \operatorname{sgn}(y_j - y_i) (\underline{x}_j - \underline{x}_i) .$$

We defined $\underline{A}_n$ as an $n \times n$ matrix:

$$\underline{A}_n = [a_{ik}] , \quad a_{ij} = - \gamma_{ij} b_{ij} , \quad \sum_{\substack{k=1 \\ k \neq i}}^{n} \gamma_{ik} b_{ik} \quad \text{as} \quad i \neq j , \quad i = j .$$

Based on the notation above, we can define the matrix $\underline{C}_n$ as:

$$\underline{C}_n = \underline{X}_c' \underline{A}_n \underline{X}_c .$$

Assumption A6

There exists a nonsingular $\underline{C}$ such that $\frac{1}{n^2} \underline{C}_n \xrightarrow{P} \underline{C}$.

This assumption is the analogue in the regular R estimates case if $\frac{1}{n} \underline{X}' \underline{X}$ converging to a positive definite matrix. We will use the above assumption in deriving our asymptotic distribution.

The following theorem will yield the projection of the negative of the gradient of

$Q^*(\underline{0})$ and hence of $\tilde{\Delta}$. We will see that the projection of $\tilde{\Delta}$ is a linear function of independent random variables. This result will greatly simplify the theoretical development of the asymptotic distribution later.

Theorem 4.5

1. $\underline{h}(\underline{Y}) = n^{-\frac{3}{2}} \sum_{i<j} (\underline{x}_j - \underline{x}_i)\, b_{ij} \operatorname{sgn}(y_j - y_i)$.

2. $\underline{h}_p(\underline{Y}) = \sum_{k=1}^{n} E_{\beta_0}(\underline{h}(\underline{Y}) \mid y_k)$.

Then

1. $\underline{h}_p(\underline{Y}) = n^{-\frac{3}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} (\underline{x}_j - \underline{x}_i)\, E_{\beta_0}[\, b_{ij} \operatorname{sgn}(y_j - y_i) \mid y_i \,]$.

2. $E[\, \underline{h}(\underline{Y}) - \underline{h}_p(\underline{Y}) \,]^2 \to 0$.

Proof.

Assume the true parameter $\underline{\beta}_0 = \underline{0}$.

Because $\underline{h}(\underline{Y}) = n^{-\frac{3}{2}} \sum_{i<j} (\underline{x}_j - \underline{x}_i)\, b_{ij} \operatorname{sgn}(y_j - y_i)$,

and let $\underline{h}_{pk} = E_{\beta_0}(\underline{h}(\underline{Y}) \mid y_k)$,

then we have $\underline{h}_{pk} = n^{-\frac{3}{2}} \sum_{i<j} (\underline{x}_j - \underline{x}_i)\, E_{\beta_0}[\, b_{ij} \operatorname{sgn}(y_j - y_i) \mid y_k \,]$.

$E_{\beta_0}[\, b_{ij} \operatorname{sgn}(y_j - y_i) \mid y_k \,]$ can be divided into following three cases:

Case 1: $k = i$

$$E_{\beta_0}[\, b_{ij} \operatorname{sgn}(y_j - y_i) \mid y_k \,] = E_{\beta_0}[\, b_{kj} \operatorname{sgn}(y_j - y_k) \mid y_k \,].$$

Case 2: $k = j$

$$E_{\beta_0}[\, b_{ij}\,\text{sgn}(\, y_j - y_i\,) \mid y_k\,] \;=\; -\; E_{\beta_0}[\, b_{ik}\,\text{sgn}(\, y_i - y_k\,) \mid y_k\,]\,.$$

Case 3: $k \neq i$ and $k \neq j$

$$E_{\beta_0}[\, b_{ij}\,\text{sgn}(\, y_j - y_i\,) \mid y_k\,]\,.$$

So $\underline{h}_{pk} = n^{-\frac{3}{2}} \{ \sum_{j>k} (\underline{x}_j - \underline{x}_k)\; E_{\beta_0}[\, b_{jk}\,\text{sgn}(\, y_j - y_k\,) \mid y_k\,] +$

$$+ \sum_{i<k} (\underline{x}_i - \underline{x}_k)\; E_{\beta_0}[\, b_{ik}\,\text{sgn}(\, y_i - y_k\,) \mid y_k\,] +$$

$$+ \sum_{\substack{i<j \\ i\neq k, j\neq k}} (\, \underline{x}_j - \underline{x}_i\,)\; E_{\beta_0}[\, b_{ij}\,\text{sgn}(\, y_j - y_i\,) \mid y_k\,] \} =$$

$$= n^{-\frac{3}{2}} \{ \sum_{j>k} (\underline{x}_j - \underline{x}_k)\; E_{\beta_0}[\, b_{jk}\,\text{sgn}(\, y_j - y_k\,) \mid y_k\,] +$$

$$+ \sum_{j<k} (\, \underline{x}_j - \underline{x}_k\,)\; E_{\beta_0}[\, b_{jk}\,\text{sgn}(\, y_j - y_k\,) \mid y_k\,] \} =$$

$$= n^{-\frac{3}{2}} \sum_{j=1}^{n} (\, \underline{x}_j - \underline{x}_k\,)\; E_{\beta_0}[\, b_{jk}\,\text{sgn}(\, y_j - y_k\,) \mid y_k\,] \}\,.$$

So $\underline{h}_p(\underline{Y}) = \sum_{k=1}^{n} \underline{h}_{pk} = n^{-\frac{3}{2}} \sum_{k=1}^{n} \sum_{j=1}^{n} (\, \underline{x}_j - \underline{x}_k\,)\, E_{\beta_0}[\, b_{jk}\,\text{sgn}(\, y_j - y_k\,) \mid y_k\,]\,.$

Next, we will show that $E(\, \underline{h}(\underline{Y}) - \underline{h}_p(\underline{Y})\,)^2 \rightarrow \underline{0}\,.$

It is easy to see that $E(\, \underline{h}(\underline{Y}) - \underline{h}_p(\underline{Y})\,) = \underline{0}\,.$

Now, we are trying to show that for all $\varepsilon > 0$, there exists an N such that for all

$n > N$, $E(\, \underline{h}_k(\underline{Y}) - \underline{h}_{p,k}(\underline{Y})\,)^2 < \varepsilon$, for all k.

Since $\underline{h}_k(\underline{Y}) - \underline{h}_{p,k}(\underline{Y}) =$

$$= n^{-\frac{3}{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} (\, x_{jk} - x_{ik}\,)\{\, \frac{1}{2}\, b_{ij}\,\text{sgn}(\, y_j - y_i\,) \;-\; E[\, b_{ij}\,\text{sgn}(\, y_j - y_i\,) \mid y_i\,]\,\} =$$

$$= n^{-\frac{3}{2}} \sum_{i<j} (x_{jk} - x_{ik})\{ b_{ij}\, sgn(y_j - y_i) - E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,] +$$

$$+ E[\, b_{ij}\, sgn(y_j - y_i) \mid y_j\,] \} .$$

Therefore $E(h_k - h_{p,k})^2 =$

$$= n^{-3} \sum_{i<j} (x_{jk} - x_{ik})^2 E\{ b_{ij}\, sgn(y_j - y_i) -$$

$$- E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,] + E[\, b_{ij}\, sgn(y_j - y_i) \mid y_j\,] \}^2 +$$

$$+ n^{-3} \sum_{\substack{i<j \\ (i,j)\neq(l,m)}} \sum_{l<m} (x_{jk} - x_{ik})(x_{mk} - x_{lk}) E\{ b_{ij}\, sgn(y_j - y_i) -$$

$$- E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,] + E[\, b_{ij}\, sgn(y_j - y_i) \mid y_j\,]\} \cdot$$

$$\cdot \{ b_{lm}\, sgn(y_m - y_l) - E[\, b_{lm}\, sgn(y_m - y_l) \mid y_l\,] + E[\, b_{lm}\, sgn(y_m - y_l) \mid y_m\,] \} .$$

It is easy to show that:

For all $\varepsilon^* > 0$, there exists N^* such that for all $n > N^*$, the 1st part $< \varepsilon^*$.

The 2nd part =

$$= n^{-3} \sum_{\substack{i<j \\ (i,j)\neq(l,m)}} \sum_{l<m} (x_{jk} - x_{ik})(x_{mk} - x_{lk}) E\{ b_{ij}\, sgn(y_j - y_i) -$$

$$- E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,] + E[\, b_{ij}\, sgn(y_j - y_i) \mid y_j\,]\} \cdot$$

$$\cdot \{ b_{lm}\, sgn(y_m - y_l) - E[\, b_{lm}\, sgn(y_m - y_l) \mid y_l\,] + E[\, b_{lm}\, sgn(y_m - y_l) \mid y_m\,] \} =$$

$$= n^{-3} \sum_{\substack{i<j \\ (i,j)\neq(l,m)}} \sum_{l<m} (x_{jk} - x_{ik})(x_{mk} - x_{lk}) E\{ b_{ij}\, sgn(y_j - y_i)\, b_{lm}\, sgn(y_m - y_l) -$$

$$- b_{ij}\, sgn(y_j - y_i)\, E[\, b_{lm}\, sgn(y_m - y_l) \mid y_l\,] +$$

$$+ b_{ij}\, sgn(y_j - y_i)\, E[\, b_{lm}\, sgn(y_m - y_l) \mid y_m\,] -$$

$$- E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,]\, b_{lm}\, sgn(y_m - y_l) +$$

$$+ E[\, b_{ij}\, sgn(y_j - y_i) \mid y_i\,]\, E[\, b_{lm}\, sgn(y_m - y_l) \mid y_l\,] -$$

$- E[\, b_{ij}\, sgn(\, y_j - y_i\,) \mid y_i\,]\, E[\, b_{lm}\, sgn(\, y_m - y_l\,) \mid y_m\,] +$

$+ E[\, b_{ij}\, sgn(\, y_j - y_i\,) \mid y_j\,]\, b_{lm}\, sgn(\, y_m - y_l\,) -$

$- E[\, b_{ij}\, sgn(\, y_j - y_i\,) \mid y_j\,]\, E[\, b_{lm}\, sgn(\, y_m - y_l\,) \mid y_l\,] +$

$+ E[\, b_{ij}\, sgn(\, y_j - y_i\,) \mid y_j\,]\, E[\, b_{lm}\, sgn(\, y_m - y_l\,) \mid y_m\,] \}$.

For the 2^{nd} part, we will divide it into two cases:

Case 1: Four indices are all different, the 2^{nd} part = 0.

Case 2: Tree indices are different.

According to Lemma 3.6, we have the following:

$b_{ij} = g_{ij}(\hat{\underline{\beta}}_0) = g_{ij}(\underline{\beta}_0) + \nabla g_{ij}'(\underline{\xi}) \cdot (\, \hat{\underline{\beta}}_0 - \underline{\beta}_0\,)$, $\underline{\xi}$ is in between $\hat{\underline{\beta}}_0$ and $\underline{\beta}_0$.

By A5, we can assume that $\frac{u_{ij}}{\sqrt{n}} = \nabla g_{ij}'(\underline{\xi}) \cdot (\, \hat{\underline{\beta}}_0 - \underline{\beta}_0\,)$.

So we only need to show that:

For all $\varepsilon^* > 0$, there exists an N^* such that for all $n > N^*$,

$$n^{-3} \sum_{\substack{i<j \\ (i,j)\neq(l,m)}} \sum_{l<m} (\, x_{jk} - x_{ik}\,)(\, x_{mk} - x_{lk}\,)\, E\{\, g_{ij}\, sgn(\, y_j - y_i\,)\, g_{lm}\, sgn(\, y_m - y_l\,) -$$

$- g_{ij}\, sgn(\, y_j - y_i\,)\, E[\, g_{lm}\, sgn(\, y_m - y_l\,) \mid y_l\,] +$

$+ g_{ij}\, sgn(\, y_j - y_i\,)\, E[\, g_{lm}\, sgn(\, y_m - y_l\,) \mid y_m\,] -$

$- E[\, g_{ij}\, sgn(\, y_j - y_i\,) \mid y_i\,]\, g_{lm}\, sgn(\, y_m - y_l\,) +$

$+ E[\, g_{ij}\, sgn(\, y_j - y_i\,) \mid y_i\,]\, E[\, g_{lm}\, sgn(\, y_m - y_l\,) \mid y_l\,] -$

$- E[\, g_{ij}\, sgn(\, y_j - y_i\,) \mid y_i\,]\, E[\, g_{lm}\, sgn(\, y_m - y_l\,) \mid y_m\,] +$

$+ E[\, g_{ij}\, sgn(\, y_j - y_i\,) \mid y_j\,]\, g_{lm}\, sgn(\, y_m - y_l\,) -$

$- E[\, g_{ij}\, sgn(\, y_j - y_i\,) \mid y_j\,]\, E[\, g_{lm}\, sgn(\, y_m - y_l\,) \mid y_l\,] +$

$$+ \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_j \,] \, E[\, g_{lm} \, sgn(y_m - y_l) \mid y_m \,] \,\} \ < \ \varepsilon^* .$$

When three indices are different with $i = l$, we have:

$$n^{-3} \sum_{i} \sum_{\substack{j>i \\ j \neq m}} \sum_{m>i} (x_{jk} - x_{ik})(x_{mk} - x_{ik}) \, E\{\, g_{ij} \, sgn(y_j - y_i) \, g_{im} \, sgn(y_m - y_i) \ -$$

$$- \ g_{ij} \, sgn(y_j - y_i) \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_i \,] \ +$$

$$+ \ g_{ij} \, sgn(y_j - y_i) \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_m \,] \ -$$

$$- \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_i \,] \, g_{im} \, sgn(y_m - y_i) \ +$$

$$+ \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_i \,] \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_i \,] \ -$$

$$- \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_i \,] \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_m \,] \ +$$

$$+ \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_j \,] \, g_{im} \, sgn(y_m - y_i) \ -$$

$$- \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_j \,] \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_i \,] \ +$$

$$+ \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_j \,] \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_m \,] \ =$$

$$= \ n^{-3} \sum_{i} \sum_{\substack{j>i \\ j \neq m}} \sum_{m>i} (x_{jk} - x_{ik})(x_{mk} - x_{ik}) \, E\{\, g_{ij} \, sgn(y_j - y_i) \, g_{im} \, sgn(y_m - y_i) \ -$$

$$- \ g_{ij} \, sgn(y_j - y_i) \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_i \,] \ -$$

$$- \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_i \,] \, g_{im} \, sgn(y_m - y_i) \ +$$

$$+ \ E[\, g_{ij} \, sgn(y_j - y_i) \mid y_i \,] \, E[\, g_{im} \, sgn(y_m - y_i) \mid y_i \,] \,\} \ =$$

$$= \ n^{-3} \sum_{i} \sum_{\substack{j>i \\ j \neq m}} \sum_{m>i} (x_{jk} - x_{ik})(x_{mk} - x_{ik}) \ \cdot$$

$$\cdot \ \{\, E[\, E(g_{ij} \, sgn(y_j - y_i) \, g_{im} \, sgn(y_m - y_i) \mid y_i) \,] \ -$$

$$- \ E[\, E[\, g_{ij} \, sgn(y_j - y_i) \, E(g_{im} \, sgn(y_m - y_i) \mid y_i) \mid y_i \,] \,] \ -$$

$$- \ E[\, E[\, E(g_{ij} \, sgn(y_j - y_i) \mid y_i) \ g_{im} \, sgn(y_m - y_i) \mid y_i \,] \,] \ +$$

$$+ \ E[\, E(g_{ij} \, sgn(y_j - y_i) \mid y_i) \ E(g_{im} \, sgn(y_m - y_i) \mid y_i) \,] \,\} \ = \ 0 .$$

The proof is similar, when tree indices are different with $i = m$, $j = l$, or $j = m$.

By the proof above, we have: for all $\varepsilon^* > 0$, three exists an N^* such that

$E(h_k(\underline{Y}) - h_{p,k}(\underline{Y}))^2 < \varepsilon^*$, for all $n > N^*$, for all k.

Hence $E(\underline{h}(\underline{Y}) - \underline{h}_p(\underline{Y}))^2 \to \underline{0}$. □

According to Theorem 4.5 above, we immediately have the following corollary.

Corollary 4.6

$$\underline{h}(\underline{Y}) - \underline{h}_p(\underline{Y}) \xrightarrow{p} \underline{0}.$$

The above theorem showed the projection of $n^{-\frac{3}{2}} S(\underline{0})$. This projection is a linear function of independent random variables. The theorem also proved that the projection has the same asymptotic distribution of the original function. This allows us to derive the asymptotic distribution based on the projection.

The following lemma is a multivariate extension of Lindeberg-Feller Theorem given by Rao(1973).

Lemma 4.7

1. $\{ \underline{U}_i \}$ are independent with means $\{ \underline{0} \}$ and covariance matrices $\{ \underline{\Sigma}_i \}$.

2. $\frac{1}{n} \sum_{i=1}^{n} \underline{\Sigma}_i \xrightarrow{p} \underline{\Sigma} \neq \underline{0}$.

3. $\frac{1}{n} \sum_{i=1}^{n} E[\|\underline{U}_i\|_2^2 \, I(\|\underline{U}_i\|_2 > \varepsilon \sqrt{n})] \to 0$, for all $\varepsilon > 0$.

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underline{U}_i \xrightarrow{d} N(\underline{0}, \underline{\Sigma}).$$

We will utilize this lemma to derive the asymptotic distribution later this chapter.

In the rest part of this chapter, we will let:

$$\underline{U}_i = \frac{1}{n}\sum_{j=1}^{n}(\underline{x}_j - \underline{x}_i)\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i).$$

Then we have:

$$\underline{h}_p(\underline{Y}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{U}_i \quad \text{and} \quad \mathrm{Var}(\underline{U}_i) = \underline{\Sigma}_i =$$

$$= \mathrm{Var}[\frac{1}{n}\sum_{j=1}^{n}(\underline{x}_j - \underline{x}_i)\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i)] =$$

$$= \frac{1}{n^2}\mathrm{Var}[\sum_{j=1}^{n}(\underline{x}_j - \underline{x}_i)\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i)] =$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}(\underline{x}_j - \underline{x}_i)(\underline{x}_k - \underline{x}_i)' \cdot$$

$$\cdot \mathrm{Cov}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i),\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)].$$

Hence $\frac{1}{n}\sum_{i=1}^{n}\underline{\Sigma}_i = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\underline{x}_j - \underline{x}_i)(\underline{x}_k - \underline{x}_i)' \cdot$

$$\cdot \mathrm{Cov}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i),\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)].$$

Because $\mathrm{Cov}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i),\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)] =$

$$= E_{\underline{\beta}_0}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i)\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)] -$$

$$- E_{\underline{\beta}_0}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i)]\ E_{\underline{\beta}_0}[\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)] =$$

$$= E_{\underline{\beta}_0}[\ E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i) \mid y_i)\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i) \mid y_i)] -$$

$$- E_{\underline{\beta}_0}(b_{ij}\,\mathrm{sgn}(y_j - y_i))\ E_{\underline{\beta}_0}(b_{ik}\,\mathrm{sgn}(y_k - y_i)) =$$

$$= E_{\beta_0}[\, E_{\beta_0}(\, b_{ij}\, \mathrm{sgn}(\, y_j - y_i\,) \mid y_i\,)\;\; E_{\beta_0}(\, b_{ik}\, \mathrm{sgn}(\, y_k - y_i\,) \mid y_i\,)\,]\,.$$

Therefore, $\frac{1}{n}\sum_{i=1}^{n} \underline{\Sigma}_i = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} (\,\underline{x}_j - \underline{x}_i\,)(\,\underline{x}_k - \underline{x}_i\,)' \cdot$

$$\cdot\; E_{\beta_0}[\, E_{\beta_0}(\, b_{ij}\, \mathrm{sgn}(\, y_j - y_i\,) \mid y_i\,)\;\; E_{\beta_0}(\, b_{ik}\, \mathrm{sgn}(\, y_k - y_i\,) \mid y_i\,)\,]\,.$$

By the discussion above, it is reasonable to make the following assumption. The assumption will be used in our theoretical proof later.

Assumption A8

Let $\underline{\Sigma}_i = \mathrm{Var}(\,\underline{U}_i\,)$, then there exists $\underline{\Sigma} \neq \underline{0}$ such that $\frac{1}{n}\sum_{i=1}^{n} \underline{\Sigma}_i \overset{p}{\to} \underline{\Sigma}$.

Based on assumption A8, the projection, $\underline{h}_p(\underline{Y})$, has the asymptotic normal distribution. It will be showed in the next theorem.

Theorem 4.8

Assume A8 holds.

Then

$\underline{h}_p(\underline{Y}) \overset{d}{\to} N(\underline{0}\,, \underline{\Sigma})\,.$

Proof.

Without loss of generality, we assume the true parameter $\underline{\beta}_0 = \underline{0}$.

Because $\underline{U}_i = \frac{1}{n}\sum_{j=1}^{n} (\,\underline{x}_j - \underline{x}_i\,)\, E_{\beta_0}(\, b_{ij}\, \mathrm{sgn}(\, y_j - y_i\,) \mid y_i\,)$,

therefore $\| \underline{U}_i \|_2^2 =$

$$= n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n} (\,\underline{x}_j - \underline{x}_i\,)'\,(\,\underline{x}_k - \underline{x}_i\,)\;\; E_{\beta_0}(\, b_{ij}\, \mathrm{sgn}(\, y_j - y_i\,) \mid y_i\,)\;\; E_{\beta_0}(\, b_{ik}\, \mathrm{sgn}(\, y_k - y_i\,) \mid y_i\,)\,.$$

Since $\| \underline{U}_i \|_2^2$ is bounded and $\frac{1}{\sqrt{n}} \| \underline{U}_i \| \to 0$,

hence, $E_{\underline{\beta}_0}[\, \| \underline{U}_i \|_2^2 \, I(\| \underline{U}_i \| > \varepsilon \sqrt{n}) \,] \to 0$, for all $\varepsilon > 0$.

By Lemma 4.7 and A8, $\underline{h}_p(\underline{Y}) \xrightarrow{d} N(\underline{0}, \underline{\Sigma})$. □

Lemma 4.9

Assume $\underline{V}_n^{p\times p} \xrightarrow{p} \underline{V}^{p\times p}$ and $\underline{U}_n^{p\times 1} \xrightarrow{d} \underline{U}^{p\times 1}$.

Then

$$\underline{V}_n^{p\times p} \, \underline{U}_n^{p\times 1} \xrightarrow{d} \underline{V}^{p\times p} \, \underline{U}^{p\times 1}.$$

Now we are ready to develop the asymptotic distribution of $\tilde{\underline{\Delta}}$.

Theorem 4.10

Assume A2 - A8 hold, and the true parameter $\underline{\beta}_0 = \underline{0}$.

Then

$$\tilde{\underline{\Delta}} \xrightarrow{d} N(0, \frac{1}{4} \underline{C}^{-1} \underline{\Sigma} \, \underline{C}^{-1}).$$

Proof.

According to Corollary 4.4 and Theorem 4.5, $\tilde{\underline{\Delta}} = \frac{1}{2} (n^2 \underline{C}_n^{-1}) \, \underline{h}(\underline{Y})$.

By Theorem 4.5, Theorem 4.8, Lemma 4.9, A2-A8, we obtain the conclusion. □

Our key result of this chapter is the asymptotic distribution theory of $\hat{\underline{\Delta}}$. As the next theorem shows, that follows the above results. The proof of the theorem below is similar to the proof in Jaeckel(1972), see Hettmansperger(1984) also.

Theorem 4.11

Assume A2 - A8 hold, and the true parameter $\underline{\beta}_0 = \underline{0}$.

Then

$$\hat{\underline{\Delta}} \overset{d}{-} N(0, \frac{1}{4} \underline{C}^{-1} \underline{\Sigma} \underline{C}^{-1}).$$

Proof.

By Theorem 4.10, for all $\delta > 0$, there exists $c > 0$ such that for sufficiently large n,

$$P(\|\tilde{\underline{\Delta}}\| \geq pc) < \frac{\delta}{2}. \quad (4.2)$$

Define $V = \min\{ Q^*(\underline{\Delta}) : \| \underline{\Delta} - \tilde{\underline{\Delta}} \| = \varepsilon \} - Q^*(\tilde{\underline{\Delta}})$, for all $\varepsilon > 0$.

Then from the definition above, $V \geq 0$.

By Theorem 4.3, we have that for all $\epsilon > 0$,

$$P(\sup_{|\underline{\Delta}| \leq pc+\varepsilon} | Q^*(\underline{\Delta}) - D^*(\underline{\Delta}) | \geq \frac{V+\epsilon}{2}) \leq \frac{\delta}{2}. \quad (4.3)$$

By (4.2) and (4.3) above, for large n, with probability greater than $1 - \delta$,

$\|\tilde{\underline{\Delta}}\| < pc$, and for all $\epsilon > 0$, for all $\|\underline{\Delta}\| \leq pc + \varepsilon$,

$$D^*(\underline{\Delta}) < Q^*(\underline{\Delta}) + \frac{V+\epsilon}{2}. \quad (4.4)$$

So for all $\underline{\Delta}$ such that $\| \underline{\Delta} - \tilde{\underline{\Delta}} \| = \varepsilon$, $\|\underline{\Delta}\| \leq \| \underline{\Delta} - \tilde{\underline{\Delta}} \| + \|\tilde{\underline{\Delta}}\| \leq \varepsilon + pc$,

and by (4.3), $| Q^*(\underline{\Delta}) - D^*(\underline{\Delta}) | < \frac{V+\epsilon}{2}$,

$$D^*(\underline{\Delta}) > Q^*(\underline{\Delta}) - \frac{V+\epsilon}{2} \geq \min\{ Q^*(\underline{\Delta}) : \| \underline{\Delta} - \tilde{\underline{\Delta}} \| = \varepsilon \} - \frac{V+\epsilon}{2} =$$

$$= V + Q^*(\tilde{\underline{\Delta}}) - \frac{V+\epsilon}{2} = Q^*(\tilde{\underline{\Delta}}) + \frac{V-\epsilon}{2} \geq D^*(\tilde{\underline{\Delta}}), \text{ by (4.4)}.$$

So for all $\underline{\Delta}$ such that $\| \underline{\Delta} - \tilde{\underline{\Delta}} \| = \varepsilon$, $D^*(\underline{\Delta}) > D^*(\tilde{\underline{\Delta}})$.

By Theorem 2.2, $D^*(\underline{\Delta})$ is a convex function.

So $D^*(\underline{\Delta}) > D^*(\tilde{\underline{\Delta}})$, for all $\underline{\Delta}$ such that $\| \underline{\Delta} - \tilde{\underline{\Delta}} \| \geq \varepsilon$.

So $D^*(\underline{\Delta}) > \min D^*(\underline{\Delta}) = D^*(\hat{\underline{\Delta}})$.

So $\| \hat{\underline{\Delta}} - \tilde{\underline{\Delta}} \| < \varepsilon$.

So for large n, with high probability, $\| \hat{\underline{\Delta}} - \tilde{\underline{\Delta}} \| < \varepsilon$.

So $\hat{\underline{\Delta}}$ and $\tilde{\underline{\Delta}}$ have the same limiting distribution. □

Next theorem, the most important theorem of this chapter, is a consequence of Theorem 4.10 and Theorem 4.11.

Theorem 4.12

Assume A1 - A8 hold, then $\sqrt{n}\,(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{d} N(0, \frac{1}{4}\underline{C}^{-1} \underline{\Sigma}\, \underline{C}^{-1})$.

Proof.

By Theorem 2.6, Theorem 4.10, and Theorem 4.11. □

Theorem 4.12 showed us that the estimate of $\underline{\beta}$ has an asymptotic normal distribution with variance-covariance matrix $\frac{1}{4}\underline{C}^{-1} \underline{\Sigma}\, \underline{C}^{-1}$. This result will enable us to develop theories of testing and confidence regions for the estimates.

CHAPTER V

ROBUST PROPERTIES OF THE ESTIMATION $\hat{\beta}$

5.1 The Breakdown Point

Let F be the cdf of ϵ_i, F_i be the cdf of y_i, and δ be the cdf of a point mass at y_0. Define $Ft = (1-t)F + t\delta$, and $\text{bias}(t) = \| \hat{\beta}(F_t) - \hat{\beta}(F) \|$, where $t = \frac{m}{n}$, n is the sample size, and m is the number of y_0's.

The following definition is given by Naranjo and Hettmansperger(1992) which is a special case of the more general definition given by Hample(1968).

<u>Definition 5.1</u>

The breakdown point $\epsilon^*_{\hat{\beta}} = \sup\{ t \le \frac{1}{2} : \text{bias}(t) < \infty \}$.

In the above definition, we have the estimates $\hat{\beta} = \arg\min D(\beta)$. If we let

$$D^{**}(\beta) = \frac{1}{n^2} D(\beta),$$

then $\hat{\beta} = \arg\min D(\beta)$ if and only if $\hat{\beta} = \arg\min D^{**}(\beta)$. Therefore, we can use $D^{**}(\beta)$ instead of $D(\beta)$ in our theoretical development.

In the following, we will use the technique that Ola Hossjer(1993) used in his paper.

Lemma 5.2

1. $a = \sup_{i,j} \{ b_{ij} \mid y_j - y_i - (\underline{x}_j - \underline{x}_i)' \underline{\beta} \mid \}.$

2. $D^{**}(\underline{\beta}) = \frac{1}{n^2} \sum_{i<j} b_{ij} \mid y_j - y_i - (\underline{x}_j - \underline{x}_i)' \underline{\beta} \mid .$

Then

1. $\frac{1}{n^2} a \le D^{**}(\underline{\beta}) \le a .$

The next is a quite mild assumption, Based on this assumption, we will develop a theorem of breakdown point for our estimates.

Assumption A9

There exists an $M > 0$ such that $\inf_{|\underline{\beta}|=1} \{ \sup_{i,j} \{ b_{ij} (\underline{x}_j - \underline{x}_i)' \underline{\beta} \} = M .$

Based on the discussions above, we will show there is a bounded estimates of $\underline{\beta}$ in the next theorem. The theorem will lead us to prove that our estimates has a high breakdown point of 50%.

Theorem 5.3

Assume A9 hold, then there exists

$$|\hat{\underline{\beta}}| < \frac{1}{M} (1 + 2n^2) \sup_{i,j} \{ b_{ij} \mid y_j - y_i \mid \}$$

such that $\hat{\underline{\beta}} = \arg\min D^{**}(\underline{\beta})$.

Proof.

According to Lemma 5.2, $D^{**}(\underline{\beta}) \ge \frac{1}{n^2} a =$

$$= \frac{1}{n^2} \sup_{i,j} \{ b_{ij} \mid y_j - y_i - (\underline{x}_j - \underline{x}_i)' \underline{\beta} \mid \} \ge$$

$\geq \frac{1}{n^2} (|\underline{\beta}| M - \sup_{i,j} \{ b_{ij} | y_j - y_i | \})$.

Because $\frac{1}{n^2} (|\underline{\beta}| M - \sup_{i,j} \{ b_{ij} | y_j - y_i | \}) \geq$

$\geq 2 \sup_{i,j} \{ b_{ij} | y_j - y_i | \}$,

when $|\underline{\beta}| \geq \frac{1}{M} (1 + 2n^2) \sup_{i,j} \{ b_{ij} | y_j - y_i | \}$.

Therefore $D^{**}(\underline{\beta}) \geq 2 \sup_{i,j} \{ b_{ij} | y_j - y_i | \}$, (5.1)

when $|\underline{\beta}| \geq \frac{1}{M} (1 + 2n^2) \sup_{i,j} \{ b_{ij} | y_j - y_i | \}$.

Also according to Lemma 5.2, $D^{**}(\underline{\beta}) \leq a = \sup_{i,j} \{ b_{ij} | y_j - y_i - (\underline{x}_j - \underline{x}_i)' \underline{\beta} | \}$.

So $D^{**}(\underline{\beta}) \leq \sup_{i,j} \{ b_{ij} | y_j - y_i | \}$. (5.2)

By Theorem 2.2, (5.1), and (5.2), we obtain the conclusion. □

According to the result of Theorem 5.3, we will show that the estimate of $\underline{\beta}$ has a high breakdown point if we choose an appropriate initial estimate.

Theorem 5.4

Assume A9 hold, and initial estimate $\hat{\underline{\beta}}_0$ has a breakdown point of $\epsilon^*_{\hat{\underline{\beta}}_0} = \frac{1}{2}$.

Then

$\hat{\underline{\beta}}$ has the same breakdown point of $\epsilon^*_{\hat{\underline{\beta}}_0}$.

Proof.

By the definition of b_{ij}, we have $\sup_{i,j} \{ b_{ij} | y_j - y_i | \} < \infty$.

So by the proof of Theorem 5.3, there exists $\hat{\beta} < \infty$, for all y_0 and for all m, where m is the number of y_0's .

So by the Definition 5.1, we have $\epsilon^*_{\hat{\beta}_0} = \min\{ \epsilon^*_{\hat{\beta}_0} , \frac{1}{2} \}$. □

5.2 The Influence Function

In the last chapter, we have proven that our estimats has a high breakdown point of 50%. The breakdown point is an important property of the estimates. It measures the globe stability of estimates to the effect of point-mass perturbations of the underlying distribution. On the other end, the influence function measures the local stability of estimates to the effect of point-mass perturbations of the underlying distribution.

Similar to the assumptions Naranjo and Hettmansperger(1994) used in their paper, we let H be the cdf of $(\underline{x}, y)$, M be the marginal cdf of $\underline{x}$, F be the conditional cdf of y given $\underline{x}$, δ_0 be the cdf of a point-mass at $(\underline{x}_0, y_0)$, and H_t be the contaminated cdf of $(\underline{x}, y)$ with $H_t = (1 - t) H + t \delta_0$.

Definition 5.5

The influence function of $\hat{\beta}$ at $(\underline{x}, y)$ is defined as:

$$IF(\hat{\beta}) = \frac{d}{dt} \hat{\beta}(H_t) |_{t=0} .$$

Based on the definition above and the discussions in the previous chapters, the influence function of our estimates $\hat{\beta}$ is bounded in both the X and Y spaces.

CHAPTER VI

IMPLEMENTATION

6.1 Parameters

In the previous chapters, we have derived the asymptotic linearity, the asymptotic distribution, and some robust properties of our estimates. In this section, we will discuss a practical issue, the choice of parameters.

As stated early, we estimate the regression coefficients $\underline{\beta}$ by minimizing the dispersion function with weights b_{ij} being defined as:

$$b_{ij} = \psi\left[\left|\frac{c\,\sigma^2\,m_i\,m_j}{z_i(\hat{\underline{\beta}}_0)\,z_j(\hat{\underline{\beta}}_0)}\right|\right],$$

where $\psi(t) = 1,\ t,\ -1,$ as $t > 1,\ -1 \le t \le 1,\ t < 1,$ $\hat{\underline{\beta}}_0$ is an initial estimate of $\underline{\beta}_0$, the true parameter. The tuning constant c is the cutoff point for outlier, and σ^2 is the variance of errors ϵ_i which rescales the residuals in the ψ function. The residual $z_i(\hat{\underline{\beta}}_0) = y_i - \underline{x}_i'\,\hat{\underline{\beta}}_0$. m_i is defined as:

$$m_i = \psi\left[\frac{b}{(\underline{x}_i - \underline{\mu}_i)'\,\underline{S}^{-1}\,(\underline{x}_i - \underline{\mu}_i)}\right].$$

$\underline{u}_i$ and $\underline{S}$ are location and covariance matrices of $\underline{X}$, respectively. b is a tuning constant. The denominator in m_i is a measure of the leverage of the i^{th} observation.

We may use the Least Median of Squares estimator (Rousseeuw & Leroy, 1987) to estimate $\underline{u}_i$ and use the Minimum Volume Ellipsoid estimator (MVE) (Rousseeuw & Leroy, 1987) to estimate $\underline{S}$.

The initial estimates $\hat{\underline{\beta}}_0$ may be estimated by using the Least Trimmed of Squares estimates (LTS) (Rousseeuw & Leroy, 1987). Recall that in the previous chapters, we assume that the initial estimate of $\underline{\beta}$ has a convergence rate of $\sqrt{n}$ and a high breakdown point of 50%. The LTS estimates meets both of the requirements.

The parameter σ^2 , which is the variance of the random errors ϵ_i , rescales the residuals in the ψ function. We can estimate σ^2 by MAD which is defined as:

$$\text{MAD} = 1.483 \ \text{med} \mid y_i - \underline{x}_i' \hat{\underline{\beta}}_0 - \text{med} (y_i - \underline{x}_i' \hat{\underline{\beta}}_0) \mid .$$

The MAD is a consistent estimate of σ^2 for normal errors, see Rousseeuw & Leroy (1987).

There are several algorithms which can be used to find a minimizing value of $D(\underline{\beta})$. In this thesis, we will use the Newton step algorithm. This algorithm uses the asymptotic quadratics of $D(\underline{\beta})$ to form the step. It is similar to the algorithm discussed by Kapenga, McKean, and Vidmar (1988).

6.2 Examples

In this section, we will discuss three data sets and several estimates which fit these data sets. We will compute the least squares estimates (LS), the Wilcoxon rand based estimates (Wilcoxon), the rank based high breakdown estimates (RHB) (the estimates proposed in this dissertation), the rank based bounded influence estimates with high breakdown weights (RBI) (Naranjo & Hettmansperger, 1990), and the least median of squares (LMS) (Rousseeuw & Leroy, 1987) for these data sets.

The first data set, the Hertzprung-Russell diagram of the star cluster CYG OB1, consists of 47 stars in the direction of Cygnus (Rousseeuw & Leroy, 1987). In this data set, the regression coefficients x is the logarithm of the effective temperature at the surface of the star, and the observations y is the logarithm of its light intensity. Several estimates for this data set are calculated and listed in Table 1 below. We also plot the data set with these fits on Figure 1 through Figure 7.

Table 1

Estimated Coefficients of the Linear Model for the Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987)

Estimates	Intercepts	Slope
LS	0.679347E+01	-0.413304E+00
Wilcoxon	0.720290E+01	-0.476636E+00
RHB	-0.608166E+01	0.252660E+01
RBI	-0.608023E+01	0.252624E+01

Figure 1 and Figure 2 below showed the least squares and the Wilcoxon rank based fits. From these two figures, we can see that both the least squares and the

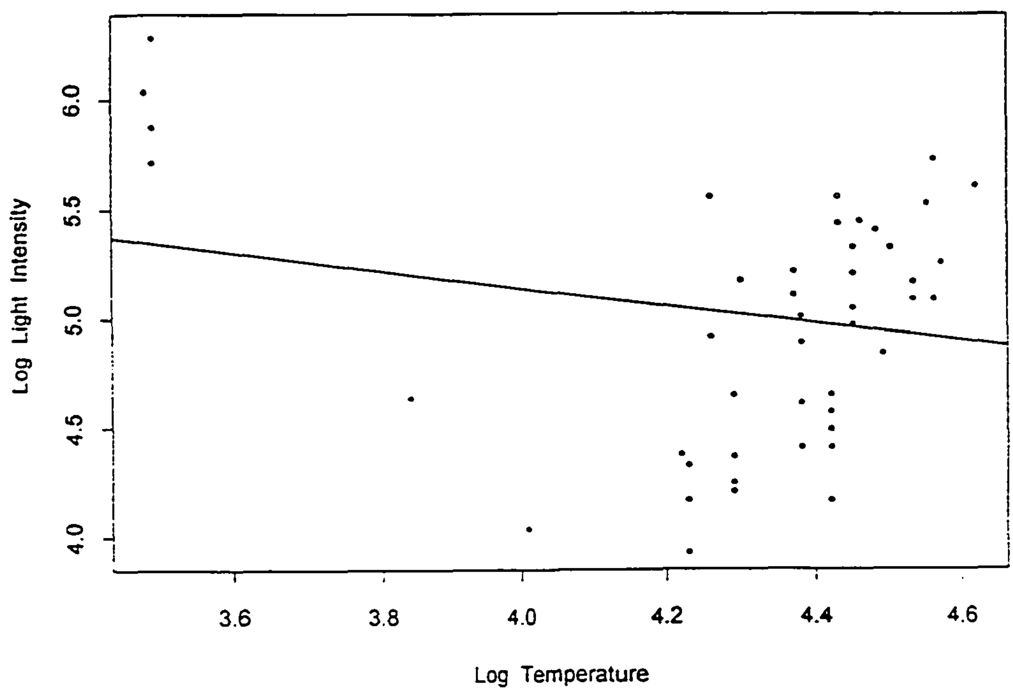

Figure 1. The Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Least Squares (LS) Fit.

Wilcoxon fits are fooled by the four outliers. They fit the data set badly. From Figure 5 through Figure 7, the fits with the RHB and the RBI are about the same. They are robust and not fooled by the outliers. The plot of the RBI fit goes through the heart of the data.

The least median of squares (LMS) and the least trimmed of squares (LTS) fits are also quite good on Figure 3, Figure 4, and Figure 7.

The second data set is a artificial data set which is generated by the model of a quadratic polynomial. The design of this model was used in a simulation study found in Naranjo, McKean, Sheather and Hettmansperger(1994). The estimates fitting this data set are listed in Table 2.

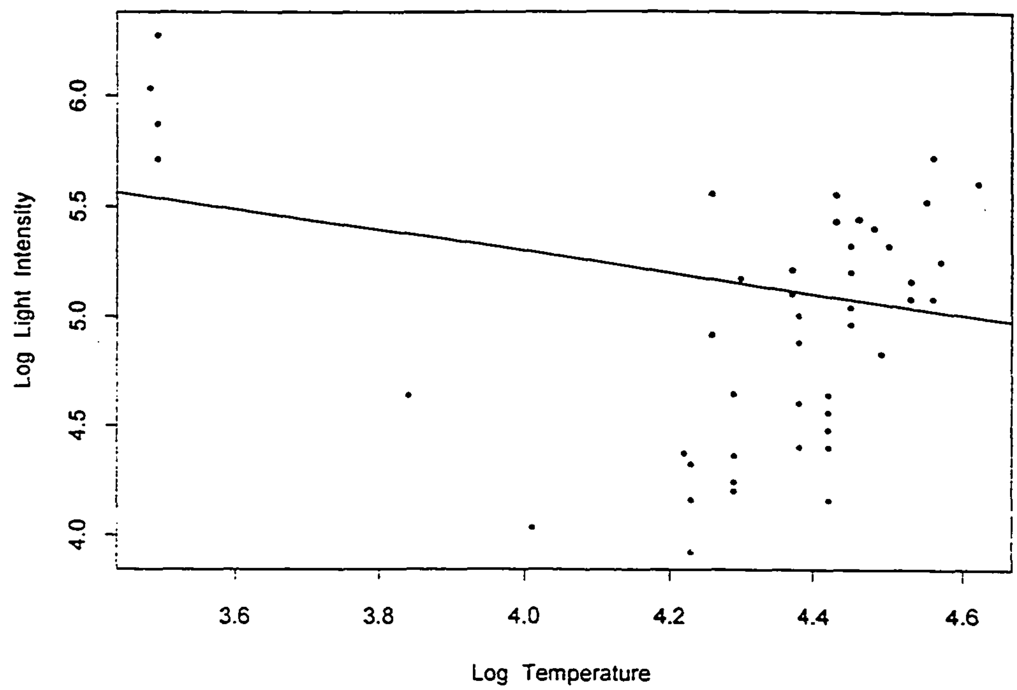

Figure 2. The Herzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Wilcoxon Rank Based Fit.

From Figure 8, we can see that the LMS fit is very poor. It misses the right side of the data set altogether. The RBI fit is also poor. It turns down long before it should. The Wilcoxon fit is quite good, as is the LS fit. The RHB fit while not as good as the Wilcoxon is an improvement over the LMS and the RBI fits. What spoils the above LMS and RBI and to a much less degree the RHB fit is that the curvature in the model is near the edge of factor space.

The last data set, the Hawkins-Bradu-Kass data set (Rousseeuw & Leroy, 1987), is an artificial data set containing 75 observations. The first 10 data points of this data set are outliers in factor space that do not follow the model. Hence, these points are bad.

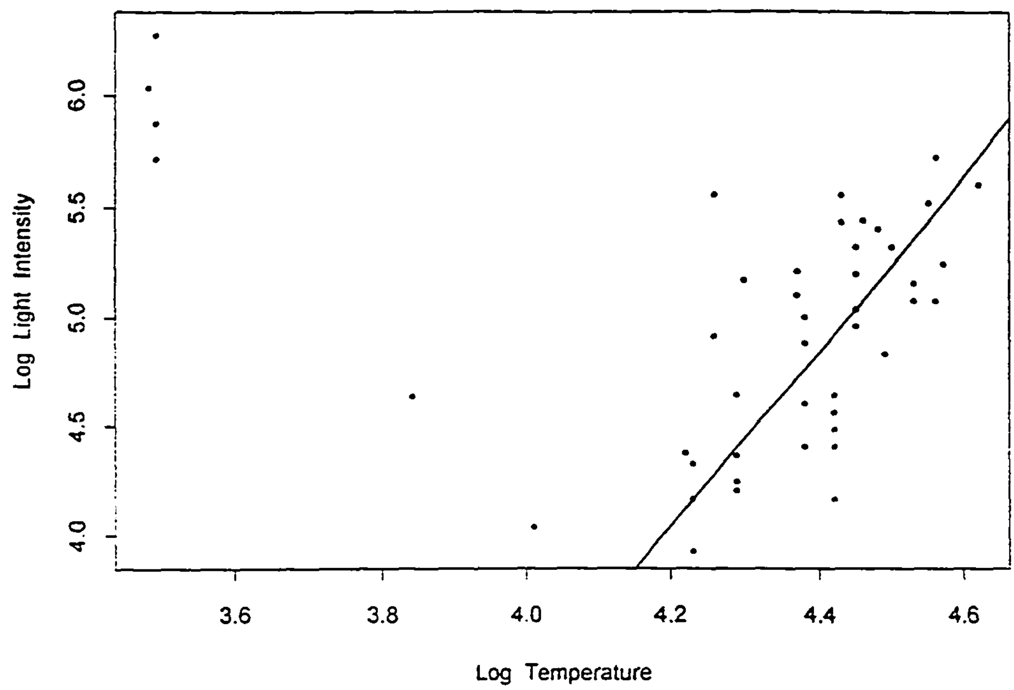

Figure 3. The Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Least Median of Squares (LMS) (Rousseeuw & Leroy, 1987) Fit.

The next four data points are outliers in factor space which do follow the model.

Several estimates fitting this data set are listed in Table 3. The Least Squares (LS) residual, the Wilcoxon residual, the GR residual, the Rank Based High Breakdown (RHB) residual, and the Rank Based Bounded Influence (RHI) residual plots for this data set will also be present in Figure 9 through Fighre 13 later this section.

From these residual plots, we can easily see that the LS fit and the Wilcoxon fit are both fooled by the outliers. They both flag the 4 good points instead of the 10 bad points. On the other hand, the RHB and the RBI fits flag the 10 outliers and show that points 11-14 follow the model.

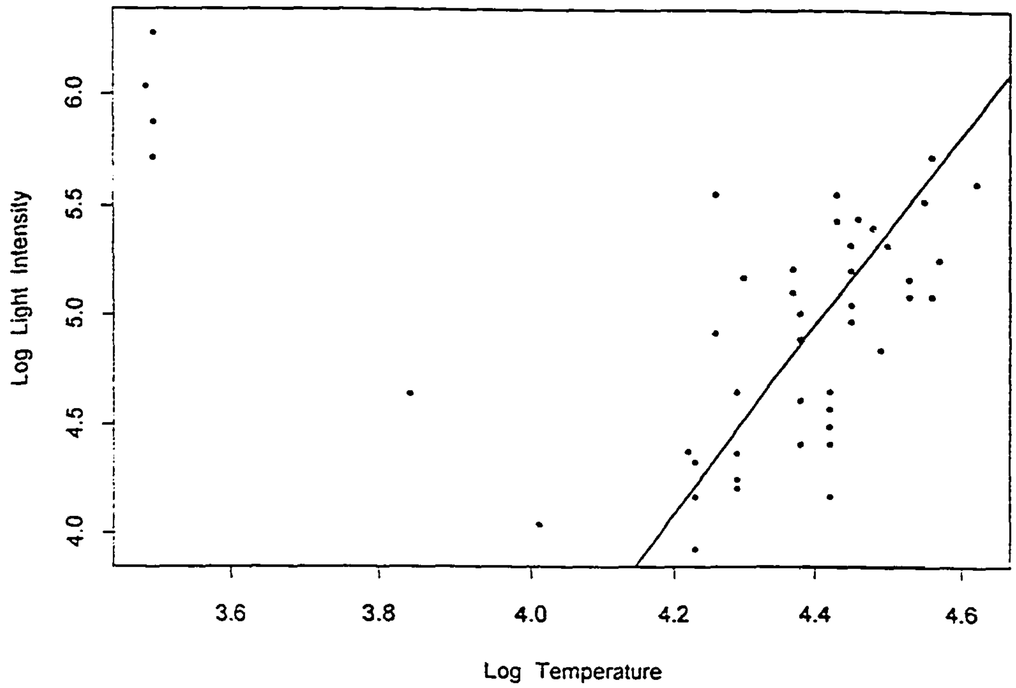

Figure 4. The Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Least Trimmed of Squares (LTS) (Rousseeuw & Leroy, 1987) Fit.

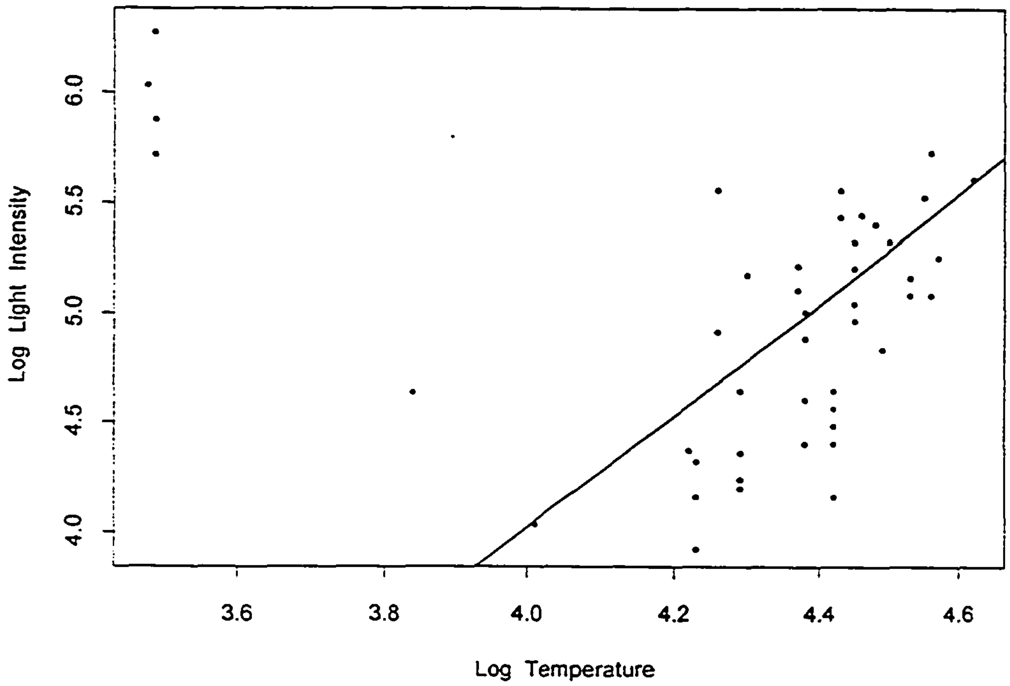

Figure 5. The Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Rank Based Bounded Influence (RHI) (Naranjo & Hettmansperger, 1990) Fit.

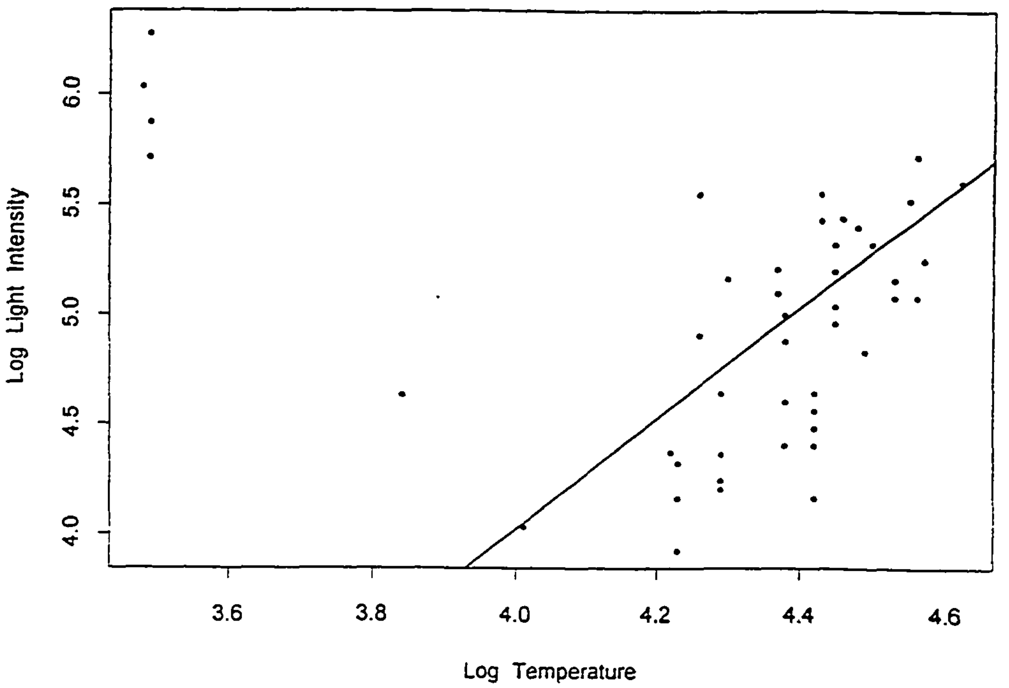

Figure 6. The Herzprung-Russell Data Set (Rousseeuw & Leroy, 1987) and the Rank Based High Breakdown (RHB) Fit.

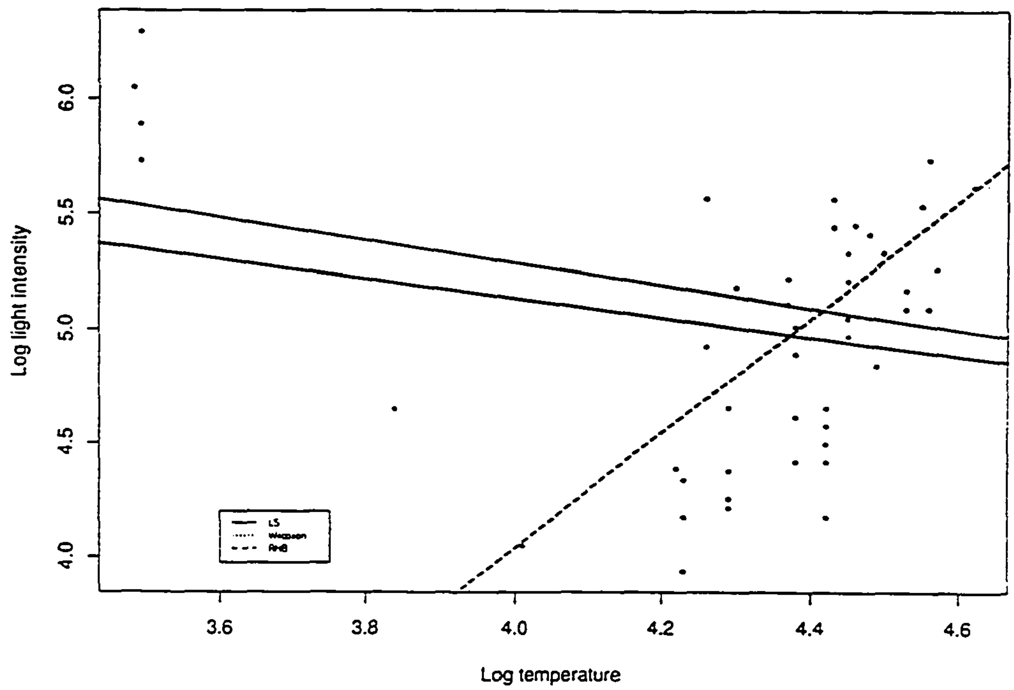

Figure 7. The Hertzprung-Russell Data Set (Rousseeuw & Leroy, 1987) With the LS, Wilcoxon, RHB, and RHI Fits.

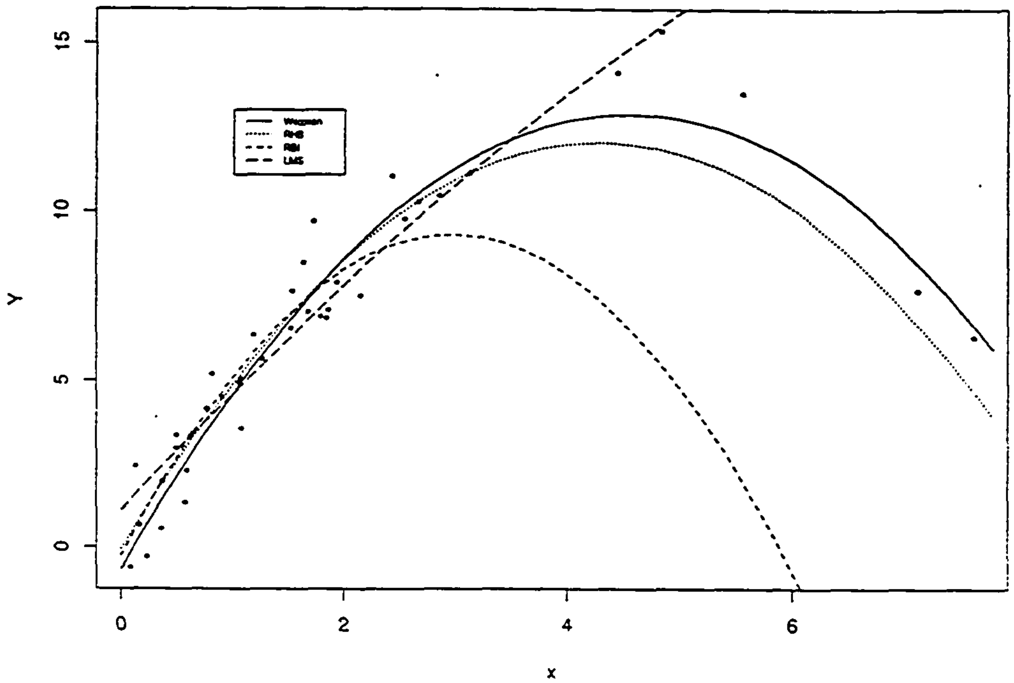

Figure 8. Run3 Data Set (Naranjo, McKean, Sheather, & Hettmansperger, 1994) With the Wilcoxon, RHB, RHI, and LMS Fits.

Table 2

Estimated Coefficients of the Linear Model for the Simulation Data Set (Naranjo et al. 1994)

Estimates	Intercept	First Degree	Second Degree
LS	-0.522891E+00	0.597353E+01	-0.655613E+00
Wilcoxon	-0.665088E+00	0.594695E+01	-0.652523E+00
RHB	-0.587292E-01	0.564280E+01	-0.658240E+00
RBI	-0.253217E+00	0.647073E+01	-0.109226E+01
LMS	-1.21	3.637	-0.137

Note:

LS - Least Squares Estimates.

Wilcoxon - Wilcoxon rank based estimates (weights identically equal to one).

RHB - Rank based high breakdown estimates (the estimates proposed in this thesis).

RBI - Rank based bounded influence estimates with high breakdown weights (Naranjo & Hettmansperger, 1990).

LMS - Least median of squares (Rousseeuw & Leroy, 1987).

Table 3

Estimated Coefficients of the Linear Model For the Hawkins Data Set (Rousseeuw & Leroy, 1987)

Estimates	Intercept	β_1	β_2	β_3
LS	-0.387551E+00	0.239186E+00	-0.334548E+00	0.383340E+00
Wilcoxon	-0.772071E+00	0.167505E+00	0.177696E-01	0.269268E+00
RHB	-0.247966E+01	0.102009E+00	0.593009E-01	-0.407682E-01
RBI	-0.177052E+00	0.827448E-01	0.525008E-01	-0.443230E-01

Note:

LS - Least Squares Estimates.

Wilcoxon - Wilcoxon rank based estimates (weights identically equal to one).

RHB - Rank based high breakdown estimates (the estimates proposed in this dissertation).

RBI - Rank based bounded influence estimates with high breakdown weights (Naranjo & Hettmansperger, 1990).

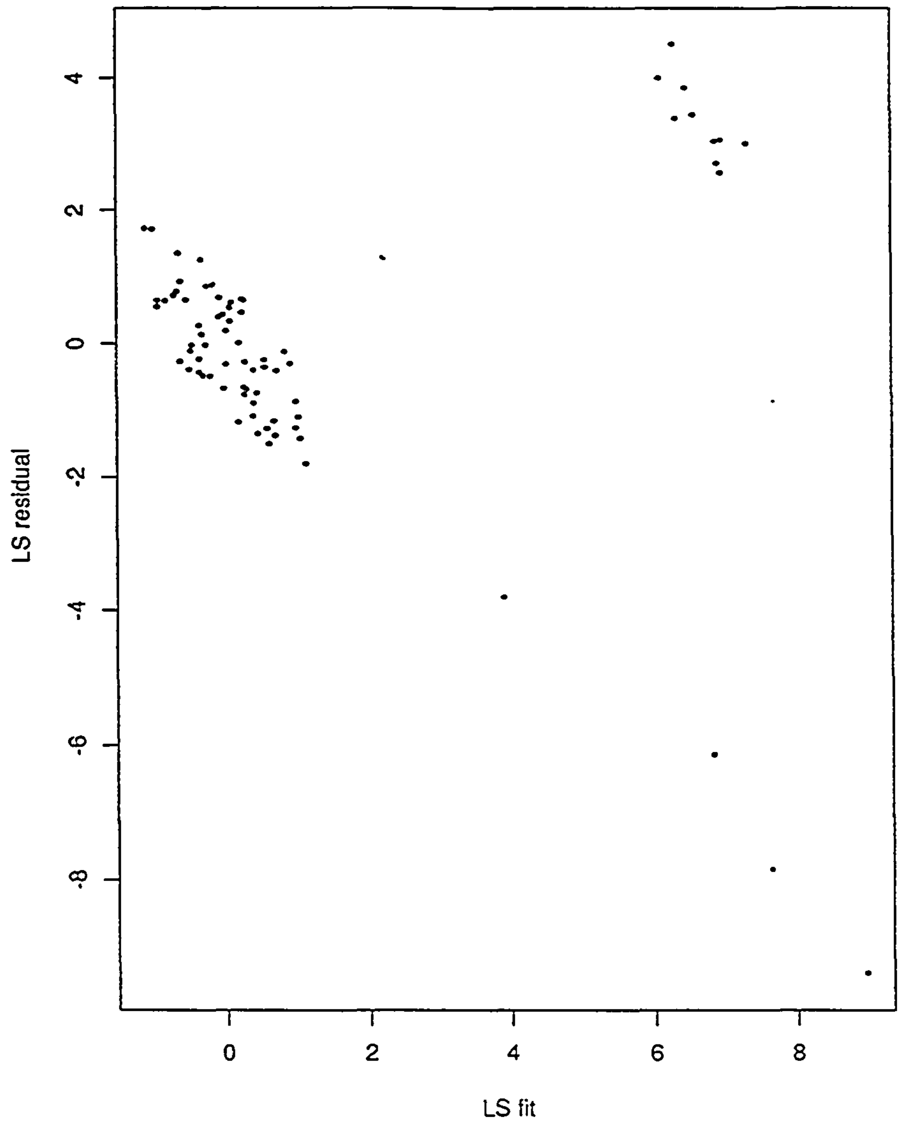

Figure 9. The Hawkins Data Set (Rousseeuw & Leroy, 1987) and the Least Squares Fit.

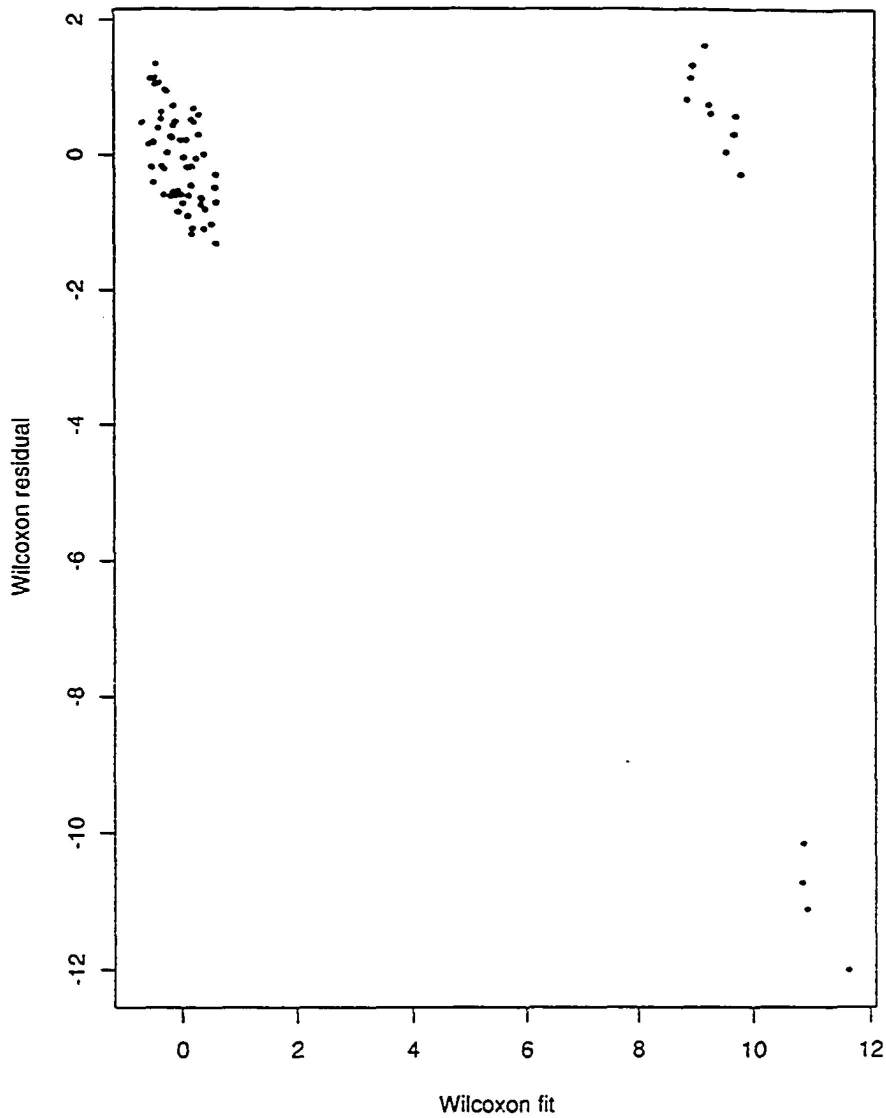

Figure 10. The Hawkins Data Set (Rousseeuw & Leroy, 1987) and the Wilcoxon Rank Based Fit.

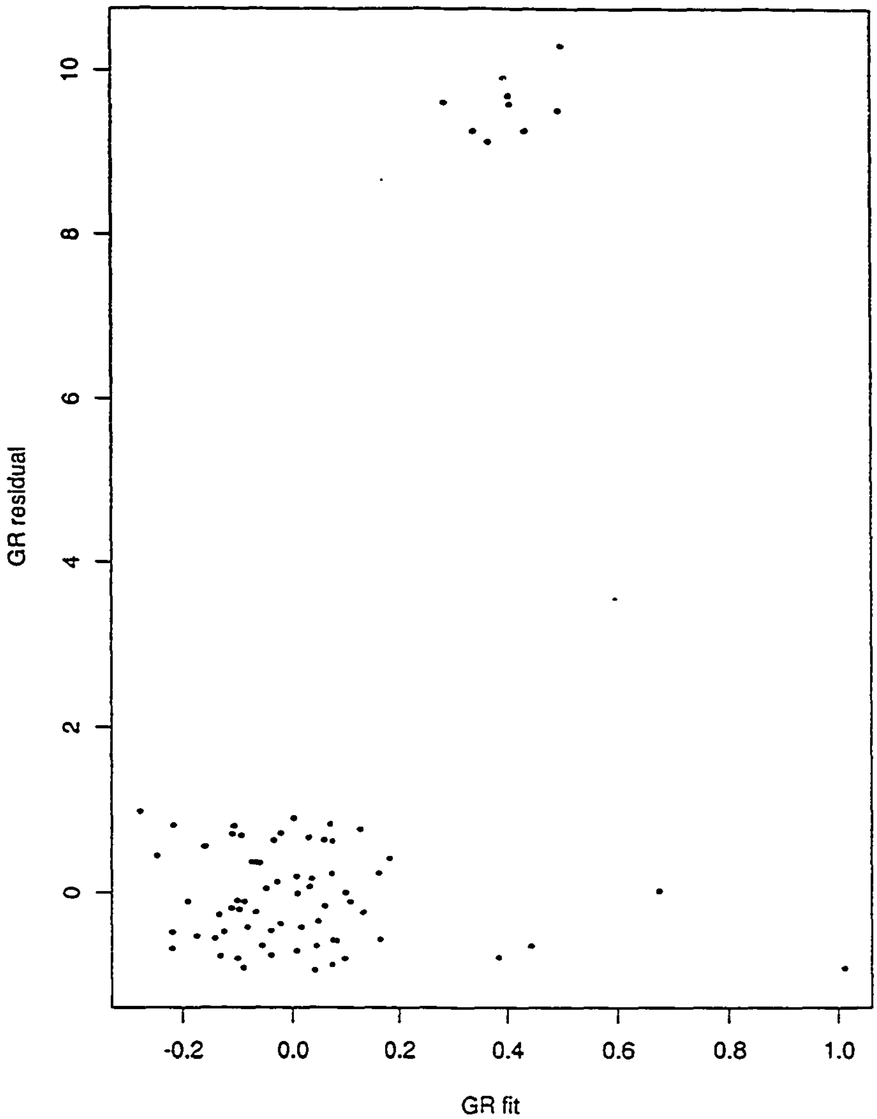

Figure 11. The Hawkins Data Set (Rousseeuw & Leroy, 1987) and the GR Fit.

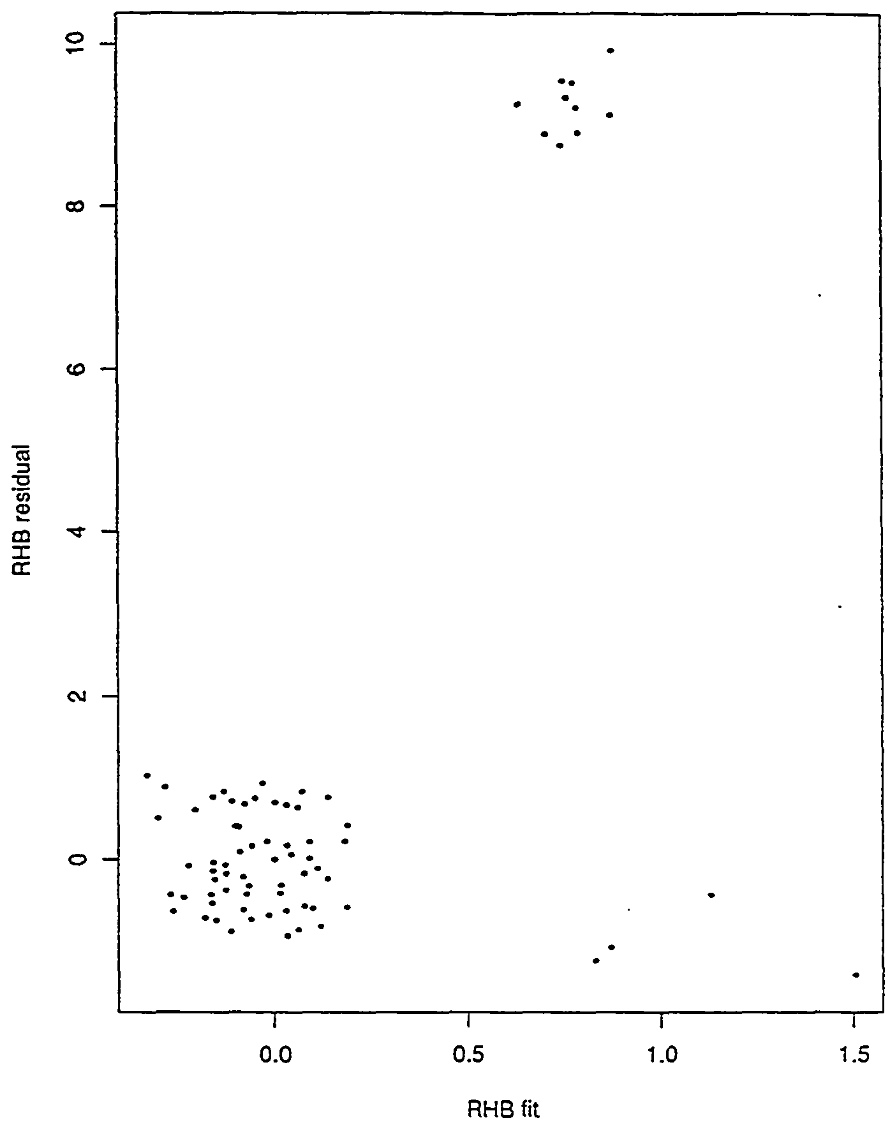

Figure 12. The Hawkins Data Set (Rousseeuw & Leroy, 1987) and the Rank Based High Breakdown (RHB) Fit.

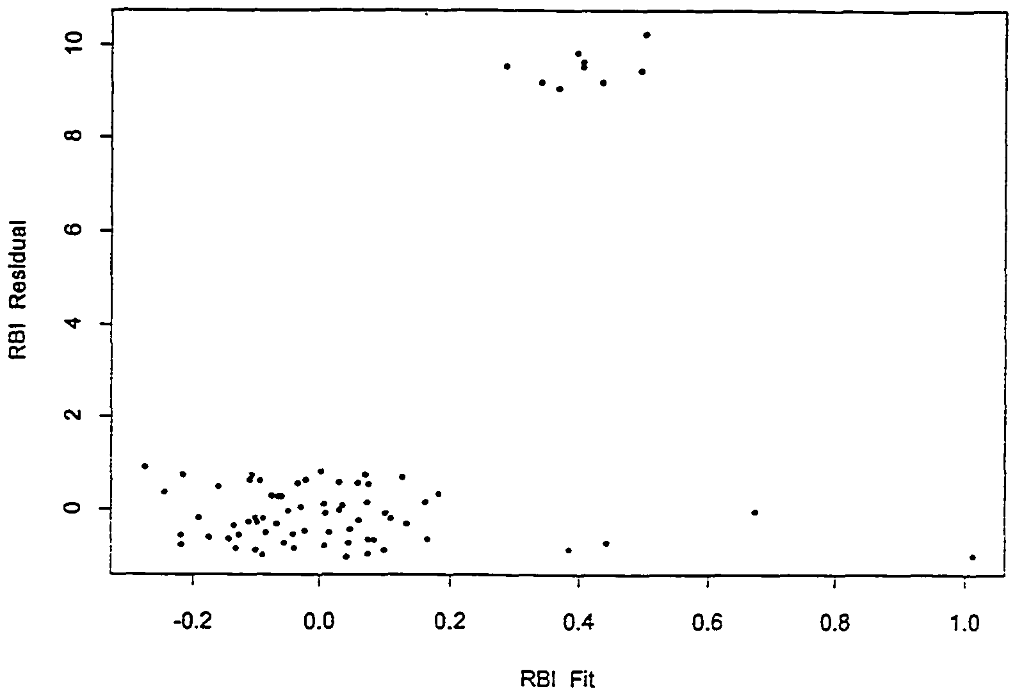

Figure 13. The Hawkins Data Set (Rousseeuw & Leroy, 1987) and the Rank Based Bounded Influence (RHI) (Naranjo & Hettmansperger, 1990) Fit.

CHAPTER VII

CONCLUSION

In this dissertation, we have discussed the linear model:

$$y_i = \underline{x}_i' \underline{\beta} + \epsilon_i, \quad i = 1, \dots, n,$$

where $\underline{x}_i'$ is the i^{th} row of $\underline{X}^{n \times p}$, ϵ_i's are independent and identically distributed with density f. We estimate the parameter $\underline{\beta}$ by minimizing the dispersion function $D(\underline{\beta})$ defined as:

$$D(\underline{\beta}) = \sum_{i<j} b_{ij} \mid z_j - z_i \mid ,$$

where $z_i = y_i - \underline{x}_i' \underline{\beta}$, $\psi(t) = 1, t, -1$, as $t > 1$, $-1 \le t \le 1$, $t < 1$, and the weight function b_{ij} are defined as:

$$b_{ij} = \psi \left[\left| \frac{c\, \sigma^2\, m_i\, m_j}{z_i(\hat{\underline{\beta}}_0)\, z_j(\hat{\underline{\beta}}_0)} \right| \right] .$$

In the previous chapters, we have derived the asymptotic linearity and asymptotic normal distribution of the estimates. The high breakdown point of 50% and the bounded influence function of the estimates have also been proved. In the last part of this thesis, we discussed three data sets with several estimates. The next step, we will try to develop theories of testing and confidence regions for the estimates. Also, we will try to find the efficiency of the estimates.

BIBLIOGRAPHY

Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics, 16, 523-531, [1.2, 2.7, 3.1, 3.6]

Apostol, T. M. (1974). Mathematical analysis, second edition. Addison-Wesley Publishing Company.

Bickel, P. J. (1975). One-Step Huber estimates in the linear model. Journal of the American Statistical Association, 70, 428-434.

Box, G. E. P. (1953). Non-normality and tests on variances, Biometrika. 40, 318-335. [1.1].

Coakley, C. W., & Hettmansperger, T. P. (1991). A bounded influence, high breakdown, efficient regression estimator.

De Jongh, P. J., De Wet, T. & Welsh, A. H. (1987). Mallows-type bounded-influence trimmed means. Journal of the American Statistical Association, 84, 805-810.

Hampel, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. in Proceedings of the Statistical Computing Section, American Statistical Association, 59-64.

Hettmansperger, T. P. (1984). Statistical inference based on ranks. John Wiley and Sons, New York.

Hettmansperger, T. & McKean, J. (1977). A robust alternative based on ranks to least squares in analyzing linear models. Technometrics, 19, 275-284.

Hossjer, Ola. (1993). Rank-based estimates in the linear model with high breakdown point. Department of Mathematics, Uppsala University, Sweden.

Huber, P. J. (1972). Robust statistics: A review. Ann. Math. Stat., 43, 1041-1067. [1.1, 3.4].

Huber, P. J. (1973). Robust regression: asymptotic, conjectures, and Monte Carlo. The Annals of Statistics, 1, 799-821.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. The Annals of Mathematical Statistics, 43, 1449-1458.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. Econometrica, 48, 1333-1346.

Krasker, W. S. & Welsch, R. E. (1982). Efficient bounded-influence regression estimation. Journal of the American Statistical Association, 77, 595-604.

Leroy, A. & Rousseeuw, P. J. (1984). Progress: A program for robust regression. Report 201, Centrum voor Statistiek en Operationazell Onderzoek, University of Brussels, Belgium.

Mallows, C. L. (1975). On some topics in robustness. Technical Memorandum, Bell Telephone Lab., Murray Hill, NJ.

McKean, J. W. & Hettmansperger, T. P. (1976). Tests of hypotheses based on ranks in the general linear model. Communications in Statistics - Theory and Methods - A, 8, 693-709.

McKean, J. W. & Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one step Estimates. Biometrika, 65, 571-579.

McKean, J. W. & Naranjo, J. D. (1990). Dianostic procedures for robust methods. Western Michigan University.

McKean, J. W. & Schrader, R. M. (1980). The geometry of robust procedures in linear models. Journal of the Royal Statistical Society Series B, 42, 366-371.

McKean, J. W. & Sheather, S. J. (1990). Small sample properties of robust analyses of linear models based on R-estimates: A survey. In press: IMA Proceedings on Robustness.

McKean, J. W., Sheather, S. J & Hettmansperger, T. P. (1990a). On the use of standardized residuals from a high breakdown gm-fit of a linear model.

McKean, J. W., Sheather, S. J. & Hettmansperger, T. P. (1990b). Regression diagnostics for rank-based methods. The Journal of the American Statistical Association.

McKean, J. W., Sheather, S. J. & Hettmansperger, T. P. (1990c). Regress diagnostics for rank-based methods II. IMA Proceedings on Robustness.

McKean, J. W. & Sievers, G. L. (1989). Rank scores suitable for analysis of linear models under asymmetric error distributions. Technometrics, 31, 207-218.

McKean, J. W., Vidmar, T. J., & Sievers, G. L. (1989). A robust two-stage multiple comparison procedure with application to a random drug screen. Biometrics, 45, 1281-1297.

Naranjo, J. D. & Hettmansperger,T. P. (1994). Bounded influence rank regression. Submitted. Journal of Royal Statistical Society, 56, 209-220.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. Biometrika, 23, 114-133. [1.1]

Plackeet, R. L. (1972). Studies in the history of probability and statistics XXIX: The discovery of the method of least squares. Biometrika, 59, 239-251. [Epigraph, 1.1, 3.4].

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871-880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. in Mathematical Statistics and Applications, Vol. B, eds. W. Grossman, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel Publishing, pp. 283-297.

Rousseeuw, P. J. & Leroy, A. M. (1987). Robust regression and Outlier Detection. Norw York: John Wiley & Sons, Inc.

Rousseeuw, P. J. & Yohai, V. (1984). Robust regression by means of S-estimators. in Robust and Nonlinear Time Series Analysis, eds. J. Franke, W. Hardle, and R. D. Martin, New York: Springer-Verlag, 256-272.

Ruppert, D., & Carrol, R. J. (1980). Trimmed least squares estimation in the linear model. Journal of the American Statistical Association, 75, 828-838.

Sievers, G.L. (1983). A weighted dispersion function for estimation in linear models. Communications in Statistics - Theory and Methods, 12(10), 1161-1179.

Simpson, D.G., Ruppert, D. & Carroll, R.J. (1992). On one-step GM estimates and stability of inferences in linear regression. Journal of the American Statistical Association, 87, 418, 439-450.

Stigler, S. M. (1981). Gauss and the invention of least squares. Ann. Stat., 9, 465-474. [1.1]

Student. (1927). Errors of routine analysis. Biometrika. 19, 151-164, [1.1].

Tukey, J. W. (1960). A survey of sampling from contaminated distributions, in Contributions to Probability and Statistics. edited by I. Olkin, Stanford University Press, Stanford, CA. [1.1].

Wainer, H. & Thissen, D. (1976). Three steps towards robust regression. Psychometrika, 41, 9-34.

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

Yohai, V. J. & Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. Journal of the American Statistical Association, 83, 406-414.