6-2002

Nonlinear Regression Based on Ranks

Ashebar Abebe
Western Michigan University

NONLINEAR REGRESSION BASED ON RANKS

by

Asheber Abebe

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 Statistics

Western Michigan University
Kalamazoo, Michigan
June 2002

NONLINEAR REGRESSION BASED ON RANKS

Asheber Abebe, Ph.D.

Western Michigan University, 2002

This study presents robust methods for estimating parameters of nonlinear regression models. The proposed methods obtain estimates by minimizing rank-based dispersions instead of the Euclidean norm. We focus on the Wilcoxon and generalized signed-rank dispersion functions. Asymptotic properties of the estimators are established under mild regularity conditions similar to those used in least squares and least absolute deviations estimation. The study also shows that by considering the generalized signed-rank dispersion we obtain a class of estimators that encompasses most of the existing popular nonlinear regression estimators. As in linear models, these rank-based procedures provide estimators that are highly efficient. This fact is further confirmed for finite samples via a simulation study. Examples illustrating the robustness of the procedure are presented.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

ACKNOWLEDGMENTS

I wish to express my gratitude to my advisor Joseph W. McKean whose continued guidance and support made this possible. My thanks go out to Gerald Sievers and Bradley Huitema with whom I had many hours of insightful discussion on subjects related to this research. Finally, I thank my family and friends for their support and encouragement.

Asheber Abebe

TABLE OF CONTENTS

Table of Contents—Continued

Table of Contents—Continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Models and Estimation

One of the most important tasks of any scientific analysis is building models to represent the relationship between the variables involved in the analysis. Often this relationship involves a response variable depending on a set of parameters in a systematic way. Most writings in the statistical literature assume that the systematic relationship is linear in the particular parameters and build models accordingly. Many interesting problems, however, are nonlinear in nature.

In this study we investigate nonlinear models where the model under consideration is suggested by the underlying mechanism which generates the data. Thus we learn of the true form of the model from the process that generates it. As mechanisms in real life are rarely deterministic, the model may depend on an unknown set of parameters, random or deterministic predictors, and random quantities which are unobservable. In other words, given the same inputs, the mechanism is unable to produce exactly the same output sequence in repeated runs.

Our interest lies in estimating the parameters upon which the model depends. Naturally, we want our estimator to behave consistently like the true value

of the parameter as more information about the underlying process becomes available. We further need the ability to predict the behavior of the model with some degree of certainty. Thus besides consistency, we need the asymptotic distribution of our estimator. We also want it to perform reasonably well, given the same amount of information, as compared to other competing estimators and to resist the influence of aberrant observations.

1.1.1 Linear Regression Based on Ranks

Linear regression based on ranks was first proposed by Jurečková (1971) and Jaeckel (1972). McKean and Schrader (1980) showed that these R estimates are based on minimizing a norm based on a score function. Hence, the geometry of these estimates is similar to that of least squares (LS) in the sense that one norm has been substituted for another. Unlike the Euclidean norm, the norm associated with R estimates leads to highly efficient, robust estimates. Chang et. al. (1999) extended these estimates to a class of high breakdown, bounded influence estimates.

These R estimates and the associated norm depend on the score function chosen. The two most popular score functions are the sign score function (L_1) and the Wilcoxon score function (linear score function). In simple location models, sign scores result in medians as the location estimates, while the Wilcoxon score function results in Hodges-Lehmann estimates; see the monograph by Hettmansperger

and McKean (1998) for a recent discussion. For normal errors, the sign and Wilcoxon estimates have asymptotic relative efficiencies (ARE's) (relative to LS) of 64% and 95%, respectively. Further, these efficiencies carry over to the linear model. The high efficiency of the Wilcoxon procedures relative to LS makes them attractive alternatives to LS procedures.

Another appealing estimation based on ranks involves the generalized signed-rank class of estimates. Just like the usual R estimates, this class uses an objective function which depends on the choice of a score function, φ^+. If φ^+ is monotone then the obective function is a norm and the geometry of the resulting robust analysis, (estimation, testing, and confidence procedures), is similar to that of the geometry of the traditional least squares (LS) analysis; see McKean and Schrader (1980). Generally this robust analysis is highly efficient relative to the LS analysis. Once again, for the simple location model, if Wilcoxon scores, $\varphi^+(u) = u$, are used then this estimate is the famous Hodges-Lehmann estimate while if sign scores are used, $\varphi^+(u) \equiv 1$, it is the sample median. If the monotonicity of φ^+ is relaxed then high breakdown estimates can be obtained; see Hössjer (1994). Thus the signed-rank family of robust estimates for the linear model contain estimates which range from highly efficient to those with high breakdown and they generalize traditional nonparametric procedures in the simple location problem.

1.1.2 Nonlinear Regression

Consider the following general nonlinear model,

$$y_i = f_i(\boldsymbol{\theta}_0) + \varepsilon_i. \quad i = 1, \ldots, n\,, \tag{1.1}$$

where each f_i are known real valued functions defined on a compact space $\boldsymbol{\Theta}$ and ε_i are random errors assumed to be independent and identically distributed.

In most cases, the dependence of f on i is borrowed from independent variables. In such cases we write

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \varepsilon_i. \quad i = 1, \ldots, n\,. \tag{1.2}$$

where $\mathbf{x}_i \in \mathcal{X} \subset \Re^q$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $f : \mathcal{X} \times \boldsymbol{\Theta} \rightarrow \Re$. The dimensions of $\boldsymbol{\Theta}$ and $\mathcal{X}$ are not necessarily the same except in the situation when $f(\mathbf{x}_i, \boldsymbol{\theta}_0) = \mathbf{x}_i^T \boldsymbol{\theta}_0$, the linear model.

We start by giving a definition of our estimation criterion, the dispersion function. Let the residual vector, $\mathbf{r}(\boldsymbol{\theta})$, be the $n \times 1$ vector whose ith element is $y_i - f_i(\boldsymbol{\theta})$.

Definition 1.1.1. A **dispersion function** is a function $D(\cdot)$ of $\mathbf{r}(\boldsymbol{\theta})$ satisfying

$$D(a\mathbf{r}(\boldsymbol{\theta}) + b\mathbf{1}) = |a| D(\mathbf{r}(\boldsymbol{\theta}))\,,$$

where $a, b \in \Re$ and $\mathbf{1}$ is the $n \times 1$ vector of ones.

The estimate of $\boldsymbol{\theta}_0$ is the argument which minimizes the dispersion function. In most instances the dispersion function is a measure of distance between

the vectors $(y_1, \cdots, y_n)^T$ and $(f_1(\boldsymbol{\theta}), \cdots, f_n(\boldsymbol{\theta}))^T$. For instance, the LS dispersion function is the Euclidean distance of the vector of residuals from the origin of $\Re^n$. Hereafter we shall write $D(\boldsymbol{\theta})$ for $D(\mathbf{r}(\boldsymbol{\theta}))$ for notational convenience.

When Θ is a subspace of the Euclidean space, $\Re^p$, the compactness assumption is equivalent to assuming that Θ is closed and bounded. In practical situations this is almost always true due to constraints imposed by the underlying mechanism. The compactness of the parameter space ensures the existence of minimizers of continuous dispersion functions as shown by Jennrich (1969). We also note that if Θ is a separable, completely regular topological space, the results continue to hold as discussed in Richardson and Bhattacharyya (1986). By the Stone-Čech Theorem, every completely regular space is a dense subspace of a compact Hausdorff space so that every continuous function defined on the original space has a continuous extension defined on the compact Hausdorff space (see Willard (1970)). Thus there is no loss of generality in assuming that the parameter space is compact.

A more restrictive assumption is to assume that Θ is a discrete subspace of $\Re^p$. Of course, this is a stronger assumption than the compactness of Θ. This path was taken by Wu (1981), who was able to give stronger results than the ones given by Jennrich (1969) in the analysis of the LS estimator of $\boldsymbol{\theta}_0$.

Let $\mathbf{f}(\boldsymbol{\theta})$ denote the $n \times 1$ vector with $f_i(\boldsymbol{\theta})$ as its ith element. Further, let

$\mathbf{y}$ denote the $n \times 1$ vector of responses. Whenever ε_i have mean zero, the surface

$$\mathcal{S}(\Theta) = \{\mathbf{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subset \Re^n ,$$

is known as the expectation surface (St. Laurent and Cook (1993)). Note that if the dimension of Θ is $p \leq n$, $\mathcal{S}$ is a p-dimensional surface in $\Re^n$. Since $\mathcal{S}$ is generally curved, most minimization problems we deal with $\mathcal{S}$ will not involve a dispersion function that is convex. Our goal is to determine a neighborhood in which $\mathcal{S}$ acts like a hyperplane so that we can achieve local convexity. If $\mathcal{S}$ is a topological n-manifold with boundary, the existence of local Euclidean spaces is immediate. Generally, if the surface $\mathcal{S}$ is smooth enough, the tangent plane at point $(\boldsymbol{\theta}, \mathcal{S}(\boldsymbol{\theta}))$ gives a good local linear approximation of $\mathcal{S}$. Note that when $\Theta \subset \Re^p$ and $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta}$ for some known $n \times p$ design matrix, $\mathbf{X}$, the expectation surface is a bounded p-dimensional hyperplane in $\Re^n$.

The existence of suboptimal minima of dispersion functions is related to the shape of the expectation surface, $\mathcal{S}$, and the distance of the response vector $\mathbf{y} \equiv (y_1, \cdots, y_n)^T$ from $\mathcal{S}$ as discussed in Pronzato and Walter (2001). The tangent approximation will not be good if the intrinsic curvature of $\mathcal{S}$ is high. This is just the ratio of the size of the quadratic term to the size of the linear term in a quadratic Taylor series approximation of $\mathbf{f}(\boldsymbol{\theta})$.

Another related concept is the concept of identifiability. In linear regression, $\boldsymbol{\theta}$ is not identifiable if $\mathbf{X}^T\mathbf{X}$ does not have full rank. In nonlinear regression non-identifiability occurs when two distinct points in Θ correspond to one point

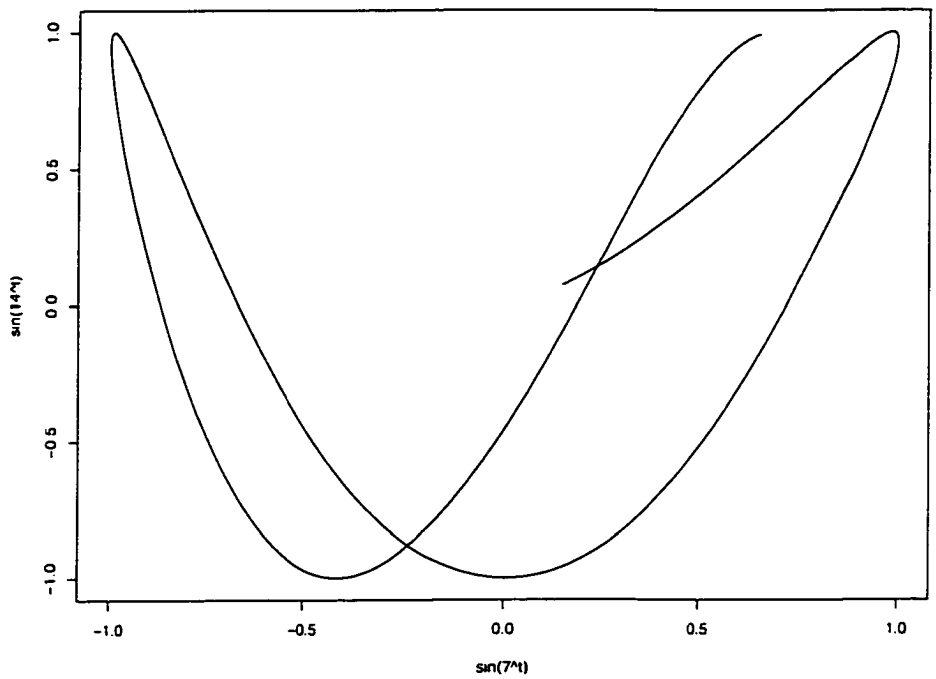

Figure 1. Expectation Surface of $\sin((7i)^{\theta})$

in $\mathcal{S}$. The geometric interpretation of this is that the space $\mathcal{S}$ folds as much as to intersect itself. In such a case the point of intersection will not be identified. As discussed in Seber and Wild (1989), the usual indicator of non-identifiability in nonlinear regression is the singularity of $(\nabla \mathbf{f})^T(\nabla \mathbf{f})$, where $\nabla \mathbf{f}$ is the $n \times p$ Jacobian matrix of $\mathbf{f}$. In our analysis we will give measure-theoretic definitions of identifiability and require that $\mathbf{f}$ is identifiable.

As the following example shows, non-identifiability could occur in quite trivial cases. Consider the model $f_i(\theta) = \sin((7i)^{\theta})$, for $i = 1, 2$ where $\theta \in [-1, 1]$. The expectation surface is actually a one dimensional curve in the two dimensional

space. Since $\sin(7^{\theta}) - \sin(14^{\theta}) = 0$ has two solutions in $[-1, 1]$, there will be no way of identifying which particular value of θ is mapped to the particular point in the expectaion surface at the roots. Figure 1 gives the expectation surface.

There has been considerable work on LS estimation of $\boldsymbol{\theta}_0$. The asymptotic properties and conditions needed for the numerical stability of the LS estimation procedure were investigated in Jennrich (1969). Manilvaud (1970) and Wu (1981) have further investigated large sample properties of the LS estimator. LS estimation in nonlinear models is a direct extension of its estimation in linear models. The same norm (Euclidean) is minimized to obtain the LS estimate of $\boldsymbol{\theta}_0$; that is, the geometry stays the same in moving from linear to nonlinear models.

Oberhofer (1982) gives sufficient conditions for the consistency of L_1 estimates of nonlinear regression parameters. We will strengthen the result by following an entirely different approach. The asymptotic normality of L_1 estimators was given by Wang (1995) under smoothness and differentiability assumptions. Even though the L_1 estimate is robust against outliers, it lacks efficiency as compared to LS estimates. Just as in the linear model its ARE is 64% relative to LS. Discussion and references concerning nonlinear estimation based on L_p norm may be found in Gonin and Money (1985).

Another approach, taken by Stromberg (1995), is to minimize the median of the square of the residuals to obtain the estimate of $\boldsymbol{\theta}_0$. This method, known as the least median of squares (LMS), was originally proposed by Rousseeuw (1984). As

Hettmansperger and Sheather (1992) show, in the linear model this method gives estimates that are unstable. This problem persists when we move to nonlinear models.

Based on geometry, R estimates can naturally be extended to nonlinear models. As in the case of LS estimates, the same norm can be used to obtain R estimates for nonlinear models as for linear models. Thus the linear model interpretation of the estimates carries over to nonlinear models.

1.2 A Motivating Example

The example we consider is a nonlinear model defined on the unit interval $(0, 1)$. This example discloses the degeneracy of the LS estimator in the presence of aberrant observations and emphasizes the need for robust estimators.

Consider a model generated using $n = 100$ (x, y) pairs satisfying

$$y = \exp x.$$

where x is a random sample taken from a uniform$(0, 1)$ distribution. Then the (x, y) values were rounded to three significant digits. The model we are fitting is a hybrid exponential model given by

$$y = \beta_0 + \exp(\beta_1 x) + e,$$

where e are random errors attributed to rounding. We expect these errors to have a random scatter since they are obtained by a symmetric rounding process. Note that the true values of the coefficients are $\beta_0 = 0$ and $\beta_1 = 1$.

As seen in Figure 2, panel (a), (b) LS gives a good fit to the data with random scatter for the residuals. Now suppose an outlier is introduced in the response space by displacing one of the points by -0.15 in the vertical direction. The effect of this outlier on the fit is not visibly apparent in Figure 2 (c), but the residual plot for LS in panel (d) shows its poor fit.

Table 1

LS Estimates for the Hybrid Exponential Model

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	RSS
Original	1.874×10^{-5}	9.999×10^{-1}	3.590×10^{-5}
With outlier	-2.373×10^{-3}	1.001×10^{0}	2.218×10^{-2}

The estimates of β_0 and β_1 along with the residual sum of squares (RSS) are given in Table 1. Also note the large increase in the RSS due to the outllier introduced. This is visibly apparent from the change in the vertical scale of the residual plot.

This example illustrates that, quite generally, LS gives estimates that are very sensitive to points that deviate from the form of the model.

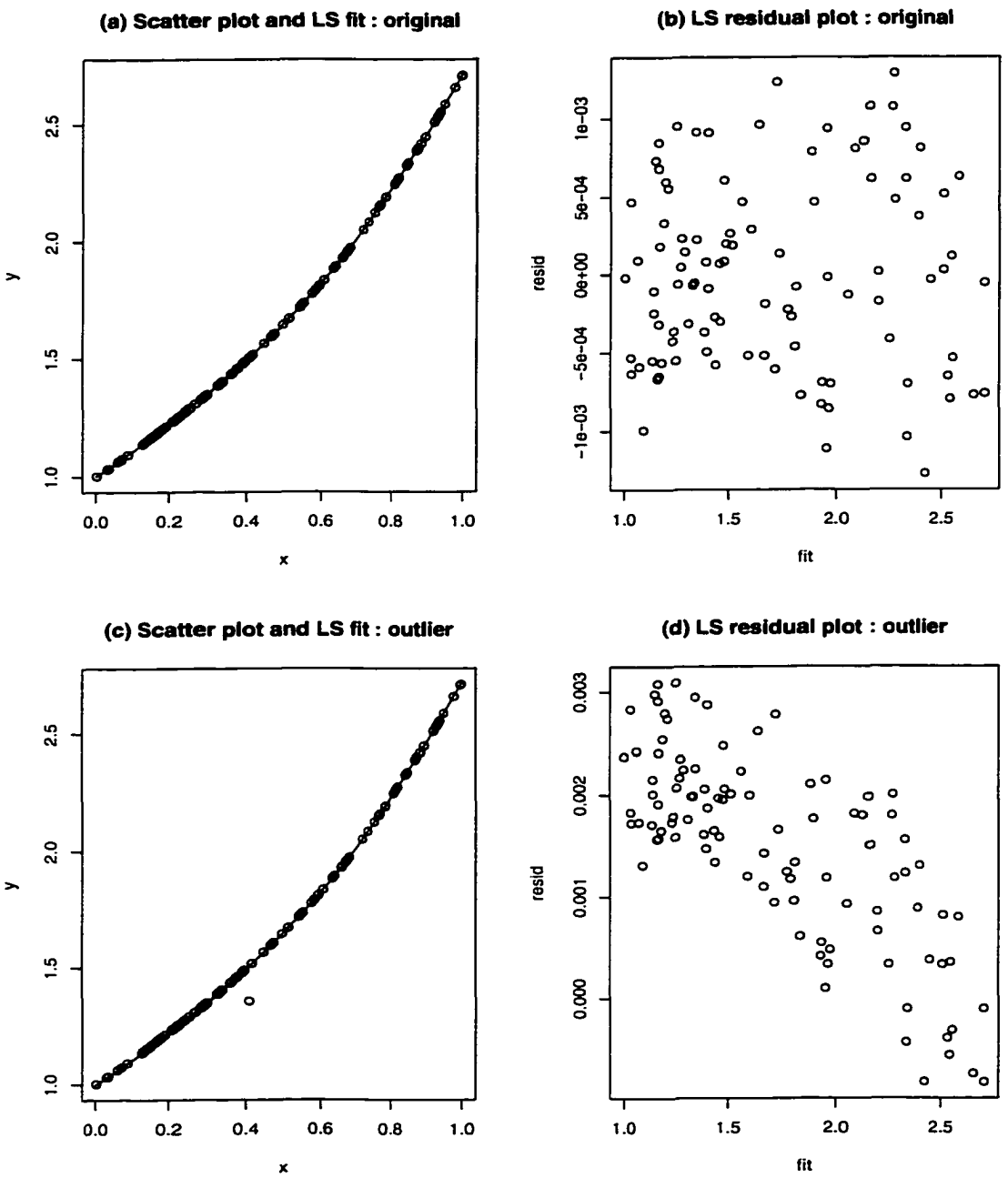

Figure 2. LS Analysis of the Hybrid Exponential Model

CHAPTER II

SOME ASYMPTOTIC RESULTS

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We shall write $h(\boldsymbol{\theta})$ instead of $h(\boldsymbol{\theta})(\omega)$ where $\omega \in \Omega$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ when there is no confusion regarding the stochastic nature of the function h (i.e. $h(\boldsymbol{\theta})$ is P-measurable). In the remainder of this study we let A° represent the interior of the space A and ∂A represent the boundary of A. The set subtraction of B from A is denoted by $A \setminus B$.

2.1 Consistency

Lemma 2.1.1 is a generalization of the results of Oberhofer (1982) and Bhattacharyya et. al. (1992), where it is used in establishing the consistency of L_1 estimators of nonlinear regression parameters.

Lemma 2.1.1. *Let $\{\Gamma_n : n \geq 1\}$ be a real valued sequence of functions defined on $\Omega \times \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is a compact space. Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$ and $\boldsymbol{\Theta}^*$ be an arbitrary compact subset of $\boldsymbol{\Theta} \setminus \{\boldsymbol{\theta}_0\}$. If*

(i) *Γ_n is uniformly continuous on $\boldsymbol{\Theta}$ for each $\omega \in \Omega$, uniformly in n.*

(ii) *there exist a sequence of real valued functions μ_n defined on $\boldsymbol{\Theta}$ such that for each $\omega \in \Omega$, $\Gamma_n(\boldsymbol{\theta}) - \mu_n(\boldsymbol{\theta}) \to 0$ in probability for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as $n \to \infty$, and*

(iii) *there exist a* $\beta = \beta(\Theta^*) > 0$ *and a* $n_0 = n_0(\Theta^*)$ *such that for all* $n \geq n_0$,

$$\inf_{\theta \in \Theta^*} \mu_n(\theta) \geq \beta .$$

then

$$\lim_{n \to \infty} P[\inf_{\theta \in \Theta^*} \Gamma_n(\theta) > 0] = 1 .$$

Proof. Let θ^* be an arbitrary point in Θ^*. Since by **(iii)** of the theorem we have a $\beta > 0$ and a n_0 such that

$$\inf_{\theta \in \Theta^*} \mu_n(\theta) \geq \beta .$$

whenever $n \geq n_0$, **(ii)** implies that

$$\lim_{n \to \infty} P[\Gamma_n(\theta^*) \geq \beta/2] = 1 . \tag{2.1}$$

Because Γ_n is uniformly continuous in θ on Θ, uniformly in n, there exist an open set K^* and a $n^* \geq n_0$ such that for $\theta^* \in K^*$ and for all $\theta \in K^*$, $|\Gamma_n(\theta) - \Gamma_n(\theta^*)| < \beta/4$ for $n \geq n^*$. Hence, by (2.1), with high probability,

$$\Gamma_n(\theta) > \beta/4 .$$

for all $\theta \in K^*$ and sufficiently large n. So, $\inf_{K^*} \Gamma_n(\theta) \geq \beta/4$ with high probability, i.e.,

$$\lim_{n \to \infty} P[\inf_{K^*} \Gamma_n(\theta) \geq \beta/4] = 1 . \tag{2.2}$$

This is true for all θ^* in Θ^*. Since Θ^* is compact, this produces a finite subcover of such sets $(K_1^*, \cdots, K_J^*)$ covering Θ^*. Therefore, K^* in (2.2) can be replaced by Θ^*. □

In the discussions of Battacharyya et. al. (1992) and Oberhofer (1982) the function μ is assumed to be the expectation of Γ_n and independent of n. Lemma 2.1.1 requires the existence of a function, μ_n so that the stochastic function $\Gamma_n - \mu_n$ converges pointwise to 0 and it makes no assumption concerning the existence of the expectation of Γ_n.

The above lemma plays a very important role in establishing the consistency of minimizers of dispersion functions. If $D_n(\boldsymbol{\theta})$ is a dispersion function, **(i) - (iii)** of Lemma 2.1.1 for $\Gamma_n(\boldsymbol{\theta}) \equiv D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)$ establish the weak consistency of the minimizer of D_n as the following lemma of Wu (1981) shows. The proof is similar to the proof of Lemma 1 of Wu (1981).

Lemma 2.1.2. *Assume*

$$\widehat{\boldsymbol{\theta}}_n \equiv \operatorname*{Argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} D_n(\boldsymbol{\theta})$$

exists. Suppose, for any arbitrary compact subset $\boldsymbol{\Theta}^*$ *of* $\boldsymbol{\Theta} \setminus \{\boldsymbol{\theta}_0\}$,

$$\liminf_{n \to \infty} \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \left[D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0) \right] > 0 \,,$$

a.s. (or in probability). Then, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ *a.s. (or in probability) as* $n \to \infty$.

Proof. We prove the a.s. convergence. Convergence in probability follows in a similar manner. Note that if $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s. is not true, then there exists $\boldsymbol{\Theta}^* \subset \boldsymbol{\Theta} \setminus \{\boldsymbol{\theta}_0\}$ such that

$$P\left[\liminf_{n \to \infty} \left\{ \widehat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}^* \right\} \right] > 0 \,,$$

which implies that

$$P\Big[\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \left[D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\right] \leq 0\Big] > 0 .$$

The result follows by taking the contrapositive of the implication. □

Lemma 2.1.2 is very general in that it only requires the existence of the limit infimum of the process as opposed to the limit which we have assumed in Lemma 2.1.1. When the limit exists and is finite, the result of Lemma 2.1.1 is equivalent to the sufficient condition given in Lemma 2.1.2 via an application of the Dominated Continuity of Measure Theorem (see, for example, Fristedt and Gray (1997)).

2.2 Asymptotic Distance Between Minimizers

The following result concerns the asymptotic distance between minimizers of dispersion functions. A version of the result was used by Jaeckel (1972). The version given here is the one found in Hjort and Pollard (1993) generalized to metric spaces.

Assume that $(\boldsymbol{\Theta}, \rho)$ is a compact metric space. Let A_n be a real valued convex random function defined on $\Omega \times \boldsymbol{\Theta}$ and let B_n be an approximation of A_n in some compact subspace, $\tilde{\boldsymbol{\Theta}}$, of $\boldsymbol{\Theta}$. For $\omega \in \Omega$, we assume that the minimizer, $\beta_n(\omega)$, of B_n is unique on $\tilde{\boldsymbol{\Theta}}$; however, we make no assumption as to the uniqueness of the minimizer, $\alpha_n(\omega)$, of A_n. Furthermore, let $C_n(\eta) = \{\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}} : \rho(\boldsymbol{\theta}, \beta_n) \leq \eta\}$.

Lemma 2.2.1. *For $\boldsymbol{\theta} \in \tilde{\Theta}$ and $\eta > 0$.*

$$P\left[\rho(\alpha_n, \beta_n) \geq \eta\right] \leq P\left[\Delta_n(\eta) \geq h_n(\eta)\right] .$$

where

$$\Delta_n(\eta) = \sup_{\boldsymbol{\theta} \in C_n(\eta)} |A_n(\boldsymbol{\theta}) - B_n(\boldsymbol{\theta})| ,$$

and

$$h_n(\eta) = \inf_{\boldsymbol{\theta} \in \partial C_n(\eta)} B_n(\boldsymbol{\theta}) - B_n(\beta_n) .$$

Proof. Let $\boldsymbol{\theta}$ be an arbitrary point outside of $C_n(\eta)$ and $\boldsymbol{\theta}^*$ be any point on $\partial C_n(\eta)$. By the convexity of A_n,

$$A_n(\boldsymbol{\theta}^*) \leq \left\{1 - \frac{\eta}{\rho(\boldsymbol{\theta}, \beta_n)}\right\} A_n(\beta_n) + \left\{\frac{\eta}{\rho(\boldsymbol{\theta}, \beta_n)}\right\} A_n(\boldsymbol{\theta}) .$$

This implies

$$\begin{aligned}
&\frac{\eta}{\rho(\boldsymbol{\theta}, \beta_n)} \left\{A_n(\boldsymbol{\theta}) - A_n(\beta_n)\right\} \\
&\qquad \geq A_n(\boldsymbol{\theta}^*) - A_n(\beta_n) \\
&\qquad = \left\{B_n(\boldsymbol{\theta}^*) - B_n(\beta_n)\right\} - \left\{A_n(\boldsymbol{\theta}^*) - B_n(\boldsymbol{\theta}^*) + A_n(\beta_n) - B_n(\beta_n)\right\} \\
&\qquad \geq \inf_{\boldsymbol{\theta} \in \partial C_n(\eta)} \left\{B_n(\boldsymbol{\theta}) - B_n(\beta_n)\right\} - 2 \sup_{\boldsymbol{\theta} \in C_n(\eta)} |A_n(\boldsymbol{\theta}) - B_n(\boldsymbol{\theta})| \\
&\qquad = h_n(\eta) - 2\Delta_n(\eta) .
\end{aligned}$$

So, if $\Delta_n(\eta) < \frac{1}{2} h_n(\eta)$, then $A_n(\boldsymbol{\theta}) > A_n(\beta_n)$ for all $\boldsymbol{\theta}$ outside $C_n(\eta)$. Thus the minimizer of A_n, α_n, has to be inside the ball $C_n(\eta)$. □

Definition 2.2.1. Two estimators, a_n and b_n, are said to be **asymptotically equivalent** if and only if $\sqrt{n}(a_n - b_n) \to 0$ in probability.

To show the asymptotic equivalence of the two minimizers, α_n and β_n, we may apply Theorem 2.2.1 with $\eta = \delta/\sqrt{n}$, $\delta > 0$. Sufficient conditions are $h_n(\delta/\sqrt{n})$ is stochastically bounded above zero and $\Delta_n(\delta/\sqrt{n})$ converges to zero in probability as $n \to \infty$.

This just says the minimizer of B_n is unique in a shrinking ball as $n \to \infty$. The process A_n is allowed to have a flat bottom as in most rank dispersion functions. In most cases, A_n is taken to be the local convex approximation of B_n via a Taylor series expansion. As we shall see in the chapters to follow, this particular technique is instrumental in establishing the asymptotic normality of our estimators.

2.3 Conditions for Normality

Noether's condition is one of the sufficient conditions for asymptotic normality of an estimator. It is given by

$$\max_{1 \le i \le n} \mathbf{x}_i^T \Big(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T \Big) \mathbf{x}_i \to 0$$

as $n \to \infty$, where $\mathbf{x}_i$ are $p \times 1$ vectors, $i = 1, \cdots, n$. The result of this section gives sufficient conditions needed for Noether's condition. The following lemma along with a proof can be found in Wu (1981).

Lemma 2.3.1. *Let* $\mathbf{x}_i, i = 1, \cdots, n$, *be* $p \times 1$ *vectors such that there exist* $\alpha_n \uparrow \infty$ *and* $\lim_{n\to\infty} \alpha_{n-1}/\alpha_n = 1$ *with* $\alpha_n^{-1} \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T$ *converging to a positive definite*

matrix Σ. *Then*

$$\max_{1\le i\le n} \mathbf{x}_i^T \Big(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T\Big)\mathbf{x}_i \;\to\; 0$$

as $n \to \infty$.

Define $\mathbf{X}$ to be the $n \times p$ matrix with the ith row given by $\mathbf{x}_i^T$. Now $H_n = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix onto the column space of $\mathbf{X}$. Another condition often used in proving asymptotic normality (known as Huber's condition) is

$$\lim_{n\to\infty} \max_{1\le i\le n} h_{iin} = 0 \,.$$

where h_{iin} is the ith diagonal entry of H_n. The following lemma given by Hettmansperger and McKean (1998) shows that Huber's condition is sufficient for Noether's condition.

Lemma 2.3.2.

$$\max_{1\le i\le n} \mathbf{x}_i^T \Big(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T\Big)\mathbf{x}_i \le \max_{1\le i\le n} h_{iin} \,.$$

2.4 Linear Combinations of Functions of Order Statistics

In our study we will deal with linear combinations of functions of order statistics (LCFOS), and therefore it is important to investigate their large sample properties. Strong laws of large numbers for LCFOS are given by Wellner (1977), Helmers (1977), and Sen (1978). A result that includes all the aforementioned works is given by van Zwet (1980). We shall discuss the work of van Zwet (1980) here.

The following definition can be found in Doob (1994). A parallel definition for random variables is given in Pollard (2002).

Definition 2.4.1. If f is a measurable function from a measure space into $\Re$. **the essential supremum of** f. denoted by **ess sup**. is the supremum of constants c for which $\{f \geq c\}$ is nonempty.

Following Doob (1994) we denote by L^p. $1 \leq p \leq \infty$. the space of measurable functions $h : (0.1) \to \Re$ for which $|h|^p$ is integrable for $1 \leq p < \infty$ and the space of essentially bounded measurable functions for $p = \infty$. The L^p norm of h is $\|h\|_p \equiv \{\int |h|^p\}^{1/p}$ for $1 \leq p < \infty$ and $\|h\|_\infty \equiv$ ess sup $|h|$ for $p = \infty$. All integrals are with respect to Lebesgue measure on (0.1).

Let $\xi_{(1)}, \ldots, \xi_{(n)}$ be order statistics from a sample of n i.i.d. uniform(0.1) random variables. Let $J_n : (0.1) \to \Re$. $n = 1.2\ldots$ be Lebesgue measurable functions and let $g : (0.1) \to \Re$ be a Borel measurable function. Define $g_n(t) \equiv g(\xi_{([nt]+1)})$.

The following lemma along with a proof can be found in van Zwet (1980).

Lemma 2.4.1. *Let* $1 \leq p \leq \infty$. $1/p + 1/q = 1$. *and suppose that* $J_n \in L^p$ *for* $n = 1.2\ldots$. $g \in L^q$, *and there exists a function* $J \in L^p$ *such that* $\lim_{n\to\infty} \int_0^t J_n = \int_0^t J$ *for all* $t \in (0.1)$. *If either*

(i) $1 < p \leq \infty$ *and* $\sup_n \|J_n\|_p < \infty$, *or*

(ii) $p = 1$ *and* $\{J_n : n = 1.2, \ldots\}$ *is uniformly integrable,*

then $\int J_n g_n \xrightarrow{a.s} \int Jg$.

This lemma is used in Chapter IV to show the strong consistency of the generalized signed-rank estimator. The idea used will be the probability integral transform which says that the distribution function has a uniform distibution on the interval $(0, 1)$. Thus order statistics from any distribution may be written in terms of order statistics from a uniform $(0, 1)$ distribution.

We now give a geometric interpretation of Lemma 2.4.1. Consider the type of convergence of $J_n \in L^p$ to $J \in L^p$ given by $\lim_{n\to\infty} \int J_n g_n = \int Jg$ for $g \in L^q$. As discussed in van Zwet (1980), **(i)** and **(ii)** of Lemma 2.4.1 are the necessary and sufficient conditions needed for the set $\{J_n, n = 1, 2, \cdots\} \subset L^p$ to be sequentially relatively compact. A set is said to be sequentially relatively compact if every sequence in the set has a subsequence that converges. Thus proving the convergence of a sequence can be done by proving its sequential relative compactness and that every convergent subsequence has the same limit point (see Fristedt and Gray (1997)).

CHAPTER III

WILCOXON ESTIMATION

3.1 Definition and Existence

Consider the nonlinear model (1.1) and let $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_n(\boldsymbol{\theta}))^T$. Given a norm $\|\cdot\|$ on n-space, a natural estimator of $\boldsymbol{\theta}$ is a value $\widehat{\boldsymbol{\theta}}$ which minimizes the distance between the response vector $\mathbf{y}$ and $\mathbf{f}(\boldsymbol{\theta})$: that is,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Argmin}} \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\| \ . \tag{3.1}$$

If the norm is the Euclidean norm then $\widehat{\boldsymbol{\theta}}$ is the LS estimate.

In this chapter, we consider the Wilcoxon norm given by,

$$\|\mathbf{u}\|_W \equiv (2n(n+1))^{-1} \sum_{i<j} |u_i - u_j| \ , \tag{3.2}$$

where $\mathbf{u}$ is a point in $\Re^n$. The quantity given in (3.2) may be represented as a linear function of the order statistics of $\mathbf{u}$ (see, for example, Hettmansperger and McKean (1998) page 73) as,

$$\|\mathbf{u}\|_W \equiv n^{-1} \sum_{i=1}^{n} a_{W,n}(i) u_{(i)} \ , \tag{3.3}$$

where $a_{W,n}(i) = \varphi_W(i/(n+1))$ where $\varphi_W(u)$ is the Wilcoxon (linear) score function given by $\varphi_W(u) = u - 1/2$. This representation is the one considered by Jaeckel (1972) in estimating linear regression parameters.

Technically the function (3.2) is a pseudo-norm; that is, it satisfies all the properties of a norm except, the property

$$\|\mathbf{u}\|_W = 0 \Leftrightarrow \mathbf{u} = \mathbf{0} .$$

is replaced by

$$\|\mathbf{u}\|_W = 0 \Leftrightarrow u_1 = u_2 = \cdots = u_n .$$

We define the Wilcoxon estimator of $\boldsymbol{\theta}_0$, denoted hereafter by $\widehat{\boldsymbol{\theta}}_{W,n}$, as

$$\widehat{\boldsymbol{\theta}}_{W,n} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Argmin}} \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|_W . \tag{3.4}$$

It will be convenient to use the notation, $D_n(\mathbf{y}, \boldsymbol{\theta}) \equiv \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|_W$. When there is no confusion we will drop the $\mathbf{y}$ from the notation.

The following lemma can be found on page 270 of Prakasa Rao (1987). The proof is credited to Landers (1968) and Strasser (1973).

Lemma 3.1.1. *Let $(\mathcal{Y}, \mathcal{A})$ be a measure space and $(\boldsymbol{\Theta}, \rho)$ be a locally compact space with countable base. Let D be a nonnegative function on $\mathcal{Y} \times \boldsymbol{\Theta}$ such that*

(i) *$D(\mathbf{y}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for all $\mathbf{y} \in \mathcal{Y}$.*

(ii) *$D(\mathbf{y}, \boldsymbol{\theta})$ is $\mathcal{A}$-measurable for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and*

(iii) *for all $\mathbf{y} \in \mathcal{Y}$ and for all $\delta > 0$, there exists a compact set $C_{\mathbf{y},\delta} \subset \boldsymbol{\Theta}$ such that*

$$\inf\{D(\mathbf{y}, \boldsymbol{\theta}) | \boldsymbol{\theta} \notin C_{\mathbf{y},\delta}\} > \delta .$$

Then there exists a measurable map $h : \mathcal{Y} \rightarrow \Theta$ *such that*

$$D(\mathbf{y}, h(\mathbf{y})) = \inf_{\boldsymbol{\theta} \in \Theta} D(\mathbf{y}, \boldsymbol{\theta}) .$$

As the next theorem shows, the following assumption suffices for the existence of the Wilcoxon nonlinear estimate:

A1: For all i, $f_i(\boldsymbol{\theta})$ is defined and continuous for all $\boldsymbol{\theta} \in \Theta$.

Theorem 3.1.1. *Under Model* (1.1) *and Assumption A1,* $\widehat{\boldsymbol{\theta}}_{W,n}$ *exists.*

Proof. Because Θ is compact it is a locally compact space with with a countable base. Parts **(i)** and **(ii)** of Lemma 3.1.1 follow trivially since under A1, by Theorem 1 of Jaeckel (1972), $D_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. Part **(iii)** of Lemma 3.1.1 follows from the fact that $D_n(\boldsymbol{\theta})$ is a nonnegative (again by Theorem 1 of Jaeckel (1972)), continuous function since for any $\delta > 0$ we can define $C_{\mathbf{y},\delta}$ to be

$$C_{\mathbf{y},\delta} = \left\{ \boldsymbol{\theta} \in \Theta : D_n(\boldsymbol{\theta}) \geq \delta + \frac{1}{2} \left| \sup_{\boldsymbol{\theta} \in \Theta} D_n(\boldsymbol{\theta}) - \delta \right| \right\} .$$

The existence of $\widehat{\boldsymbol{\theta}}_{W,n}$ follows from Lemma 3.1.1. □

Another popular score besides the Wilcoxon, is the sign score function given by $\varphi_S(u) = \text{sgn}(u - (1/2))$. The norm associated with this score function is (3.3) but with the sign scores $a_{S,n}(i) = \varphi_S(i/(n+1))$. Let $\widehat{\boldsymbol{\theta}}_S$ denote the estimate based on this norm. Its existence follows in the same way as the existence of the Wilcoxon. In order to see the relationship between $\widehat{\boldsymbol{\theta}}_S$ and the L_1 estimate of $\boldsymbol{\theta}$,

denote the L_1 norm by,

$$\|\mathbf{u}\|_{L_1} \equiv n^{-1} \sum_{i=1}^{n} |u_i| \,. \tag{3.5}$$

Let $\widehat{\boldsymbol{\theta}}_{L_1}$ denote the estimate based on this norm. The following lemma is the nonlinear analogue of Theorem 3.8.1 of Hettmanpserger and McKean (1998).

Lemma 3.1.2. *If $\boldsymbol{\Theta}$ is a compact subset of $\Re^p$, then*

$$\mathbf{f}(\widehat{\boldsymbol{\theta}}_{L_1}) = \mathbf{f}(\widehat{\boldsymbol{\theta}}_S) + med\{y_i - f_i(\widehat{\boldsymbol{\theta}}_S)\}\mathbf{1} \,.$$

where, $\mathbf{1}$ is a vector of n ones.

Proof. Since $\boldsymbol{\theta} \in \Re^p$, we may write it as $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$. Let $\mathbf{F}(\boldsymbol{\theta})$ be the $n \times p$ matrix with ijth element $f_i(\theta_j)$. Let $\Xi = \{\mathbf{F}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\Xi_1 = \{[\mathbf{1}\ \mathbf{F}(\boldsymbol{\theta})] : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Any two vectors, $\mathbf{v} \in \Xi$ and $\mathbf{v}_c \in \Xi_1$, are related as $\mathbf{v} = a\mathbf{1} + \mathbf{v}_c$ where $a \in \Re$. We have

$$\|\mathbf{y} - \mathbf{v}\|_{L_1} = \|\mathbf{y} - a\mathbf{1} - \mathbf{v}_c\|_{L_1} \geq \|\mathbf{y} - \text{med}\{\mathbf{y} - \mathbf{v}_c\}\mathbf{1} - \mathbf{v}_c\|_{L_1} \,,$$

with the last inequality due to the fact that the sample median minimizes the L_1 distance between a vector and the space spanned by $\mathbf{1}$. This implies, for any $\mathbf{v} \in \Xi$,

$$\|\mathbf{y} - \mathbf{v}\|_{L_1} \geq \|\mathbf{y} - \text{med}\{\mathbf{y} - \mathbf{v}_c\}\mathbf{1} - \mathbf{v}_c\|_{L_1} = \|\mathbf{y} - \mathbf{v}_c\|_S \,. \tag{3.6}$$

since $\text{sgn}(y_i - \text{med}\{\mathbf{y} - \mathbf{v}_c\} - v_{ci}) = \text{sgn}(R(y_i - v_{ci}) - (n+1)/2)$ and the sign scores sum to 0. Using the same argument we can show that

$$\|\mathbf{y} - \text{med}\{\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_S)\}\mathbf{1} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_S)\|_{L_1} = \|\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}_S)\|_S \,. \tag{3.7}$$

Putting (3.6) and (3.7) together completes the proof. □

Oberhofer (1982) obtained asymptotic theory for the L_1 estimate $\widehat{\boldsymbol{\theta}}_{L_1}$. In Chapter IV we will strengthen the results of Oberhofer (1982) using a generalized signed-rank dispersion function. Using Lemma 3.1.2, the asymptotic theory for the signed estimator $\widehat{\boldsymbol{\theta}}_S$ can easily be derived.

As in the linear case, the ARE of the L_1 estimate, relative to LS at normal errors, is low, 63%. In the next two sections, we derive the asymptotic theory (consistency and asymptotic normality) of the Wilcoxon nonlinear estimator $\widehat{\boldsymbol{\theta}}_{W,n}$. As in the linear model situation, we show that the Wilcoxon nonlinear estimate has an ARE of 95%, relative to LS at normal errors. Thus the Wilcoxon estimate provides a highly efficient, nonlinear estimate of $\boldsymbol{\theta}_0$.

3.2 Consistency

Before we establish the consistency of $\widehat{\boldsymbol{\theta}}_{W,n}$ we introduce some helpful notation. Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ be points in Θ. We denote the residuals at $\boldsymbol{\theta}$ by $e_i(\boldsymbol{\theta}) \equiv y_i - f_i(\boldsymbol{\theta})$ for $1 \leq i \leq n$. For $1 \leq i, j \leq n$, we define $W_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \equiv |e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta})| - |e_i(\boldsymbol{\theta}^*) - e_j(\boldsymbol{\theta}^*)|$. Further, let

$$h_i^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \equiv f_i(\boldsymbol{\theta}^*) - f_i(\boldsymbol{\theta}),$$

$$h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \equiv h_i^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - h_j^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*), \quad \text{and}$$

$$\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \equiv n^{-1} \sum_{i=1}^{n} \{h_i^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\}^2 .$$

Let G denote the distribution function of $\varepsilon_i - \varepsilon_j$.

We need the following assumptions:

A2: $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$.

A3: $\lim_{n\to\infty} n^{-1}\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

A4: $G(0) = 1/2$.

A5: There exist $\eta > 0$ and n_0 such that for all $n \geq n_0$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$, where $\boldsymbol{\Theta}^*$ is a closed subset of $\boldsymbol{\Theta} \setminus \{\boldsymbol{\theta}_0\}$.

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} n^{-2} \sum_{i<j} |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \times$$

$$\min\left\{G(|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2) - 1/2\ ,\ 1/2 - G(-|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2)\right\} \geq \eta\ .$$

Assumption A3 is the same as Oberhofer's (1982) assumption A4. Jennrich (1969) assumes that $\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ converges uniformly to a continuous function $\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ for all $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$ in $\boldsymbol{\Theta}$ and $\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This of course implies A3. Wu (1981) assumes that $n\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ diverges as n approaches infinity. This is weaker than Jennrich's assumption and A3 since it does not restrict $\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ to converge at the rate of n. Assuming the existence of the dot product $\sum f_i\varepsilon_i$, in light of Lemma 2.1.2, one can easily observe that Wu's condition is sufficient for the consistency of the LS estimator. This, however, is not true in our case as Δ_n only comes into play as part of an upper bound on $D_n(\boldsymbol{\theta})$. Discussion on Assumption A5 follows Theorem 3.2.1.

Now let Θ^* be a closed subset of Θ not containing $\boldsymbol{\theta}_0$. The weak consistency of $\widehat{\boldsymbol{\theta}}_{W,n}$ will then follow if for all such Θ^* and every $\boldsymbol{\theta} \in \Theta^*$,

$$\lim_{n\to\infty} P(\inf_{\boldsymbol{\theta}\in\Theta^*}[D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)] > 0) = 1 . \tag{3.8}$$

Lemma 3.2.1. *Under A3.*

$$\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} - E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} \to 0 .$$

in probability.

Proof. The statement of the lemma can be written as.

$$\lim_{n\to\infty} P\left(\left|[2n(n+1)]^{-1}\sum_{i<j}[W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) - E(W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))]\right| > \delta\right) = 0 ,$$

for all $\delta > 0$. Now applying Markov's inequality followed by Minkowski's, triangular and Jensen's inequalities (see Petrov (1995)) we get,

$$\begin{aligned}
&P\left(\left|[2n(n+1)]^{-1}\sum_{i<j}[W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) - E(W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))]\right| > \delta\right) \\
&\qquad \le [2\delta n(n+1)]^{-1}E|\sum_{i<j}[W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) - E(W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))]| \\
&\qquad \le [2\delta n(n+1)]^{-1}\sum_{i<j}E|W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) - E(W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))]| \\
&\qquad \le [2\delta n(n+1)]^{-1}\sum_{i<j}\{E|W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| + |E(W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))|\} \\
&\qquad \le [\delta n(n+1)]^{-1}\sum_{i<j}E|W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| .
\end{aligned} \tag{3.9}$$

But,

$$|W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| = \big||e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta})| - |\varepsilon_i - \varepsilon_j|\big|$$

$$\leq |h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|$$

$$\leq |h_i^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)| + |h_j^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)|.$$

This implies that,

$$[\delta n(n+1)]^{-1}\sum_{i<j} E|W_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \leq \frac{n-1}{\delta n(n+1)}\sum_{i=1}^{n}|h_i^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)|$$

$$\leq \frac{n-1}{\delta(n+1)}\left\{n^{-1}\Delta_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)\right\}^{1/2},$$

which goes to zero as $n \to \infty$ by A3. This combined with (3.9) completes the proof. □

Lemma 3.2.2. *Under A4,*

$$E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} \geq [2n(n+1)]^{-1}\sum_{i<j}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\times$$

$$\min\{G(|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|/2) - 1/2\ .\ 1/2 - G(-|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|/2)\}\ .$$

Proof. Note that,

$$D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0) = [2n(n+1)]^{-1}\sum_{i<j}\left\{|(\varepsilon_i - \varepsilon_j) + h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| - |\varepsilon_i - \varepsilon_j|\right\}.$$

It is easy to show that if T is a random variable with distribution function F_T and $F_T(0) = 1/2$, then for any constant k,

$$E(|T+k|-|T|) = 2I(k \leq 0)\int_0^{-k}\{|k|-x\}dF_T(x) + 2I(k>0)\int_{-k}^{0}\{|k|+x\}dF_T(x)\ . \tag{3.10}$$

Applying this we obtain,

$$[2n(n+1)]^{-1}\sum_{i<j}E\{|(\varepsilon_i-\varepsilon_j)+h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|-|\varepsilon_i-\varepsilon_j|\}$$
$$=[n(n+1)]^{-1}\sum_{(i,j)\in A}\int_0^{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}\{|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|-x\}dG(x)$$
$$+[n(n+1)]^{-1}\sum_{(i,j)\in B}\int_{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}^{0}\{|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|+x\}dG(x)\ .$$

where A and B are a partition of the set $\{(i,j): i<j\}$ given by $A=\{(i,j): i<j$ and $h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\le 0\}$ and $B=\{(i,j): i<j$ and $h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)>0\}$. Restricting the ranges of integration and applying A4 as in Oberhofer (1982) we get,

$$[2n(n+1)]^{-1}\sum_{i<j}E\{|(\varepsilon_i-\varepsilon_j)+h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|-|\varepsilon_i-\varepsilon_j|\}$$
$$\ge[2n(n+1)]^{-1}\sum_{(i,j)\in A}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\{G(-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/2)-1/2\}$$
$$+[2n(n+1)]^{-1}\sum_{(i,j)\in B}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\{1/2-G(-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/2)\}$$
$$\ge[2n(n+1)]^{-1}\sum_{i<j}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\times$$
$$\min\{G(|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/2|)-1/2,\ 1/2-G(-|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)/2|)\}\ .$$

□

We now state and prove the main theorem of this section.

Theorem 3.2.1. *Under A1-A5, $\widehat{\boldsymbol{\theta}}_{W,n}$ is weakly consistent for $\boldsymbol{\theta}_0$.*

Proof. Let

$$\Gamma_n(\boldsymbol{\theta})\equiv D_n(\boldsymbol{\theta})-D_n(\boldsymbol{\theta}_0)\ ,$$

and

$$\mu_n(\boldsymbol{\theta}) \equiv E\{\Gamma_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\} .$$

For any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^* \subset \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}^*$ is a closed set not containing $\boldsymbol{\theta}_0$, under A5 and Lemma 3.2.2, there exist a $\beta = \beta(\boldsymbol{\Theta}^*) > 0$ and a $n_0 = n_0(\boldsymbol{\Theta}^*)$, such that for all $n \geq n_0$,

$$\inf_{\boldsymbol{\Theta}^*} \mu_n(\boldsymbol{\theta}) \geq \beta . \tag{3.11}$$

Because

$$\Gamma_n(\boldsymbol{\theta}) \leq \frac{n-1}{n+1}\{n^{-1}\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}^{1/2},$$

it is uniformly continuous in $\boldsymbol{\theta}$ on $\boldsymbol{\Theta}$, uniformly in n. Moreover, by Lemma 3.2.1, we have $\Gamma_n(\boldsymbol{\theta}) - \mu_n(\boldsymbol{\theta}) \rightarrow 0$ in probability for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus by Lemma 2.1.1 we have that

$$\lim_{n\to\infty} P[\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \Gamma_n(\boldsymbol{\theta}) > 0] = 1 .$$

The result follows from Lemma 2.1.2. □

Assumption A5 is similar to that assumed in Oberhofer (1982). As the following lemma shows assumption A5 is an identifiability assumption. We will give the following definition of identifiability of measurable functions. A similar definition can be found in Seber and Wild (1989).

Definition 3.2.1. Let $\mathbf{f}$ be λ-measurable, where λ is a σ-finite measure. The parameters of the nonlinear regression problem $\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ are said to be **unidentifiable** if there exist two distinct points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ such that $\lambda\{\mathbf{f}(\boldsymbol{\theta}_1) \neq \mathbf{f}(\boldsymbol{\theta}_2)\} = 0$.

Lemma 3.2.3. *If for each* $\boldsymbol{\theta}$, $\mathbf{f}(\boldsymbol{\theta})$ *is a real measurable function on a measure space* $(\omega, \mathcal{F}, \lambda)$, *where* λ *is a* σ*-finite measure, then a necessary condition for A5 to hold is,*

$$\lambda\{\mathbf{f}(\boldsymbol{\theta}) \neq \mathbf{f}(\boldsymbol{\theta}_0)\} > 0 ,$$

for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Proof. Assume that there exists $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ distinct from $\boldsymbol{\theta}_0$ such that $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta}_0)$ λ-a.e. For such $\boldsymbol{\theta}$, $|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| = 0$ λ-a.e. implying that A5 fails to hold. □

If f is a known function as in (1.2) that depends on a set of random predictors, in addition to the parameters, then we obtain a natural extension of this. The following corollary gives the result.

Corollary 3.2.1. *Let* $(\Omega, \mathcal{F}, P)$ *be the underlying probability space and* $f_i(\boldsymbol{\theta}) = f(\mathbf{z}_i, \boldsymbol{\theta})$. *Assume* $\mathbf{z}_i$ *are n independent identically distributed m dimensional random vectors with range* $\mathcal{Z} \subset \Re^m$. *Then a necessary condition for A5 is that for all* $\mathbf{z}_i \in \mathcal{Z}$,

$$P(f(\mathbf{z}_i(\omega), \boldsymbol{\theta}) = f(\mathbf{z}_i(\omega), \boldsymbol{\theta}_0)) < 1 ,$$

for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *and each* $\omega \in \Omega$.

Proof. Obvious. □

The following lemma gives sufficient conditions for A5 to hold.

Lemma 3.2.4. *If*

A5.1: *$\varepsilon_i - \varepsilon_j$ have density g continuous at 0 with $g(0) > 0$, and*

A5.2: *There exist $\eta > 0$ and n_0 such that for all $n \geq n_0$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$*

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} n^{-2}\{h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}^2 \geq \eta \, .$$

then A5 is true.

Proof. Since $G(0) = 1/2$, applying a first order Taylor series expansion of G about 0, we have

$$n^{-2} \sum_{i<j} |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \times \min \left\{ G\left(\frac{|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|}{2} \right) - 1/2 \, , \, \frac{1}{2} - G\left(\frac{-|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|}{2} \right) \right\}$$

$$= n^{-2} \sum_{i<j} \{h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}^2 g(t)/2 \, ,$$

where $t \in (-|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2, |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2)$. Since g is continous there is an interval $(-\delta, \delta)$ over which $g > 0$ for some $\delta > 0$. Moreover, since h_{ij} is continuous and $\boldsymbol{\Theta}^*$ is arbitrary, the interval $(-|h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2, |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|/2)$ can be made a subset of $(-\delta, \delta)$. The result follows from A5.2. □

For most practical purposes A5.1 and A5.2 are easier to verify than A5. As discussed in the remark immediately following the definition of assumption A5, A5.2 is a double-indexed version of Jennrich's (1969) assumption on $\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

3.3 Asymptotic Normality

In this section we will investigate the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{W,n}$. The theory for the asymptotic normality uses tangent planes. A 'manifold-type' prop-

erty of the expectation surface, $\mathcal{S}$, is furnished by the consistency of $\widehat{\boldsymbol{\theta}}_{W,n}$ in the locality of $\boldsymbol{\theta}_0$. Hence the asymptotic theory of linear models plays a very important role in showing the asymptotic normality of $\widehat{\boldsymbol{\theta}}_{W,n}$. Details concerning the asymptotic properties of rank estimators of linear model parameters can be found in Hettmansperger and McKean (1998).

In addition to A1 - A5, we will assume that the following conditions are satisfied.

N1: For $i = 1, \ldots, n$, $f_i(\boldsymbol{\theta})$ is continuously differentiable at $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}$.

N2: The sequence of matrices

$$n^{-1} \sum_{i=1}^{n} \{\nabla f_i(\boldsymbol{\theta}_0)\}\{\nabla f_i(\boldsymbol{\theta}_0)\}^T$$

converges to a positive definite matrix $\Sigma(\boldsymbol{\theta}_0)$ where $\nabla f_i(\boldsymbol{\theta})$ is the $p \times 1$ derivative vector of $f_i(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

N3: The error density g has a finite Fisher information.

For $i = 1, \ldots, n$, let

$$e_i^*(\boldsymbol{\theta}) \equiv y_i - f_i(\boldsymbol{\theta}_0) - \{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \, .$$

Note that $e_i^*(\boldsymbol{\theta})$ are the error terms of the linear regression,

$$y_i^* = \mathbf{x}_i^{*T}\boldsymbol{\theta}_0 + \varepsilon_i \, , \tag{3.12}$$

where $y_i^* = y_i - f_i(\boldsymbol{\theta}_0) + \{\nabla f_i(\boldsymbol{\theta}_0)\}^T \boldsymbol{\theta}_0$ and $\mathbf{x}_i^* = \nabla f_i(\boldsymbol{\theta}_0)$. Define the corresponding Wilcoxon dispersion function as,

$$D_n^*(\boldsymbol{\theta}) \equiv [2n(n+1)]^{-1} \sum_{i<j} |e_i^*(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})| \,. \tag{3.13}$$

Furthermore, let,

$$\widetilde{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{Argmin}}\, D_n^*(\boldsymbol{\theta}) \,. \tag{3.14}$$

Under assumption N2, we apply Lemma 2.3.1 to obtain Noether's condtion.

$$\max_{1 \le i \le n} \{\nabla f_i(\boldsymbol{\theta}_0)\}^T \left[\sum_{i=1}^{n} \{\nabla f_i(\boldsymbol{\theta}_0)\} \{\nabla f_i(\boldsymbol{\theta}_0)\}^T \right]^{-1} \{\nabla f_i(\boldsymbol{\theta}_0)\} \to 0 \,, \tag{3.15}$$

as $n \to \infty$. The following theorem uses this condition to establish the asymptotic normality of $\widetilde{\boldsymbol{\theta}}_n$. A rigorous derivation of the result may be found in Hettmansperger and McKean (1998).

Theorem 3.3.1. *Under model* (3.12) *and assumptions N2, N3 we have,*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \tau_\varphi^2 \Sigma(\boldsymbol{\theta}_0)) \,, \tag{3.16}$$

where

$$\tau_\varphi^{-1} = \int (G(t) - 1/2)(-g'(t)/g(t)) dG(t) \,. \tag{3.17}$$

and $\Sigma(\boldsymbol{\theta}_0)$ *is given in assumption N2.*

Proof. The result follows from Corollary 3.5.6 of Hettmansperger and McKean (1998) via (3.15). □

Let ∇f_0 be the $n \times p$ matrix with the ith row given by $\{\nabla f_i(\boldsymbol{\theta}_0)\}^T$. The projection operator onto the tangent plane at $\boldsymbol{\theta}_0$ of the expectation surface, $\mathcal{S}$, is given by

$$P_n = \nabla f_0 (\nabla f_0^T \nabla f_0)^{-1} \nabla f_0^T .$$

Assume

H1: $\lim_{n\to\infty} \max_{1\le i\le n} p_{iin} = 0$, where p_{iin} is the ith diagonal entry of P_n.

The following corollary of Theorem 3.3.1 shows that H1 may be used to prove the asymptotic normality of $\widetilde{\boldsymbol{\theta}}_n$ whenever it is convenient.

Corollary 3.3.1. *Under model* (3.12) *and assumptions H1, N3 we have,*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \tau_\varphi^2 \Sigma(\boldsymbol{\theta}_0)) ,$$

where

$$\tau_\varphi^{-1} = \int (G(t) - 1/2)(-g'(t)/g(t)) dG(t) , \tag{3.18}$$

and $\Sigma(\boldsymbol{\theta}_0)$ *is given in assumption N2.*

Proof. The proof follows by Lemma 2.3.2 and Theorem 3.3.1. □

The approach we follow to prove the asymptotic normality is via Slutsky's Theorem (see Serfling (1980)). Recall that Lemma 2.2.1 gives a probabilistic bound on the asymptotic distance between the minimizers of A_n, a convex process with a possibly flat bottom, and B_n, a process whose minimizer is asymptotically unique in a neighborhood which shrinks at the rate of $1/\sqrt{n}$. In light

of Lemma 3.2.2 and the consistency of $\widehat{\boldsymbol{\theta}}_{W,n}$, we have a neighborhood where the minimum of D_n is unique as $n \to \infty$. Moreover, the process D_n^* is convex as shown in Theorem 1 of Jaeckel (1972). Thus, D_n and D_n^* may be treated as B_n and A_n, respectively, in Lemma 2.2.1.

The following lemma in addition to (3.16) gives the main result of this section which is given in Theorem 3.3.2 below.

Lemma 3.3.1. *Under A1 - A5, N1 - N3,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n} - \widetilde{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{P}} \mathbf{0}\,.$$

Proof. Let ρ be any metric on $\boldsymbol{\Theta}$. For $\delta > 0$, define $M_n(\delta) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \rho(\widehat{\boldsymbol{\theta}}_{W,n}, \boldsymbol{\theta}) \leq \delta/\sqrt{n}\}$. By Lemma 2.2.1, sufficient conditions for

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n} - \widetilde{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{P}} \mathbf{0}\,.$$

are

$$\sup_{\boldsymbol{\theta} \in M_n(\delta)} |D_n(\boldsymbol{\theta}) - D_n^*(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0 \quad . \quad \text{and} \tag{3.19}$$

$$\inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} \{D_n(\boldsymbol{\theta}) - D_n(\widehat{\boldsymbol{\theta}}_{W,n})\} \geq \beta > 0\,, \tag{3.20}$$

for all $\delta > 0$ and sufficiently large n.

To verify (3.19) notice that

$$\begin{aligned} |D_n(\boldsymbol{\theta}) - D_n^*(\boldsymbol{\theta})| &\leq [2n(n+1)]^{-1} \sum_{i<j} |e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta}) + e_j^*(\boldsymbol{\theta})| \\ &\leq \frac{n-1}{2n(n+1)} \sum_{i=1}^{n} |e_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta})|. \end{aligned}$$

But, for $1 \leq i \leq n$,

$$|e_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta})| \leq |f_i(\boldsymbol{\theta}_0) - f_i(\boldsymbol{\theta})| + |\{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)|$$

$$\leq |f_i(\boldsymbol{\theta}_0) - f_i(\widehat{\boldsymbol{\theta}}_{W,n})| + |f_i(\boldsymbol{\theta}) - f_i(\widehat{\boldsymbol{\theta}}_{W,n})|$$

$$+ |\{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{W,n})| + |\{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\widehat{\boldsymbol{\theta}}_{W,n} - \boldsymbol{\theta}_0)|.$$

Moreover, since f_i are uniformly continuous on Θ and $\widehat{\boldsymbol{\theta}}_{W,n}$ is weakly consistent for $\boldsymbol{\theta}_0$,

$$|f_i(\boldsymbol{\theta}_0) - f_i(\widehat{\boldsymbol{\theta}}_{W,n})| + \sup_{\boldsymbol{\theta} \in M_n(\delta)} |f_i(\boldsymbol{\theta}) - f_i(\widehat{\boldsymbol{\theta}}_{W,n})| \xrightarrow{\mathcal{P}} 0. \tag{3.21}$$

We also have $\|\nabla f_i(\boldsymbol{\theta}_0)\| < \infty$ by N1 and the compactness of Θ. This gives us.

$$\sup_{\boldsymbol{\theta} \in M_n(\delta)} |\{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{W,n})| + \sup_{\boldsymbol{\theta} \in M_n(\delta)} |\{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\widehat{\boldsymbol{\theta}}_{W,n} - \boldsymbol{\theta}_0)| \xrightarrow{\mathcal{P}} 0. \tag{3.22}$$

The expressions in (3.21) and (3.22) establish (3.19).

Proceeding to show (3.20) notice that by the definition of $\widehat{\boldsymbol{\theta}}_{W,n}$ and continuity of $D_n(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ we have,

$$\inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} \{D_n(\boldsymbol{\theta}_0) - D_n(\widehat{\boldsymbol{\theta}}_{W,n})\} \geq 0 \quad \text{a.s.} \tag{3.23}$$

We also have,

$$\inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} \{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} \geq \inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\}$$

$$+ \inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} \left[D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0) - E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\}\right].$$

But by Lemma 3.2.2 and assumption A5, there exist $\eta > 0$ and n_0 such that for all $n \geq n_0$,

$$\inf_{\boldsymbol{\theta} \in \partial M_n(\delta)} E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} \geq \eta.$$

Also by Lemma 3.2.1.

$$\inf_{\boldsymbol{\theta}\in\partial M_n(\delta)} \left[D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0) - E\{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\}\right] \xrightarrow{\mathcal{P}} \mathbf{0}.$$

Thus for sufficiently large n we have,

$$\inf_{\boldsymbol{\theta}\in\partial M_n(\delta)} \{D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)\} \geq \eta. \tag{3.24}$$

The expressions (3.23) and (3.24) give (3.20). The proof is complete. □

Theorem 3.3.2. *Under A1 - A5 and N1 - N3.*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \tau_\varphi^2 \Sigma(\boldsymbol{\theta}_0)) .$$

where τ_φ *is as given in* (3.18).

Proof. Follows immediately from Lemma 3.3.1 via an application of Slutsky's Theorem. □

As a simple corollary, a useful asymptotic representation of the Wilcoxon estimate is obtained. By (3.16), it follows, as in Hettmansperger and McKean (1998), that,

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \tau_\varphi (n^{-1}\mathbf{X}^{*T}\mathbf{X}^*)^{-1} n^{-1/2}\mathbf{X}^{*T}\{G(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}_0) - 1/2\} + o_p(1) , \tag{3.25}$$

where $\mathbf{X}^*$ is the $n \times p$ matrix with the ith row given by $\{\nabla f_i(\boldsymbol{\theta}_0)\}^T$ and $\mathbf{y}^*$ is an $n \times 1$ vector with the ith component $y_i - f_i(\boldsymbol{\theta}_0) + \{\nabla f_i(\boldsymbol{\theta}_0)\}^T\boldsymbol{\theta}_0$. Now applying Lemma 3.3.1 we have the same asymptotic representation for the Wilcoxon estimate, i.e.,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n} - \boldsymbol{\theta}_0) = \tau_\varphi (n^{-1}\mathbf{X}^{*T}\mathbf{X}^*)^{-1} n^{-1/2}\mathbf{X}^{*T}\{G(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}_0) - 1/2\} + o_p(1) . \tag{3.26}$$

Based on (3.26), we can obtain the influence function of the Wilcoxon estimate. Assume f_i depends on a set of predictors $\mathbf{z}_i \in \mathcal{Z} \subset \Re^m$ as $f_i(\boldsymbol{\theta}) = f(\mathbf{z}_i, \boldsymbol{\theta})$. Assume also that f is a continuous function of $\boldsymbol{\theta}$ for each $\mathbf{z} \in \mathcal{Z}$ and is a measurable function of $\mathbf{z}$ for each $\boldsymbol{\theta} \in \Theta$ with respect to a σ-finite measure. Under these assumptions, the representation above gives us the local influence function of the Wilcoxon estimate at the point $(\mathbf{z}_0, y_0)$.

$$\mathrm{IF}(\mathbf{z}_0, y_0; \widehat{\boldsymbol{\theta}}_{W,n}) = \tau_\varphi \{\Sigma(\boldsymbol{\theta}_0)\}^{-1} \{G(y_0) - 1/2\} \nabla f(\mathbf{z}_0, \boldsymbol{\theta}_0) .$$

Note that the influence function is unbounded if the tangent plane of $\mathcal{S}$ at $\boldsymbol{\theta}_0$ in unbounded. This phenomenon corresponds to the existence of high leverage points in linear regression. The analysis in Chapter V gives a possible remedy for this problem by considering a weighted form of the Wilcoxon norm. This parallels GR estimation in linear models given by Naranjo and Hettmansperger (1994).

3.4 Estimation Algorithm

There are several ways of estimating $\widehat{\boldsymbol{\theta}}_{W,n}$. The algorithm we use is based on the representation given in (3.26) which will play an important role in establishing a numerical procedure to estimate $\widehat{\boldsymbol{\theta}}_{W,n}$. Let .

$$\mathbf{h}_n(\boldsymbol{\theta}) = \boldsymbol{\theta} + \tau_\varphi \Big(\{\mathbf{X}^*(\boldsymbol{\theta})\}^T \{\mathbf{X}^*(\boldsymbol{\theta})\} \Big)^{-1} \{\mathbf{X}^*(\boldsymbol{\theta})\}^T \Big\{ G\big(\mathbf{y}^*(\boldsymbol{\theta}) - \{\mathbf{X}^*(\boldsymbol{\theta})\}\boldsymbol{\theta}\big) - 1/2 \Big\} . \tag{3.27}$$

where $\mathbf{X}^*(\boldsymbol{\theta})$ is the $n \times p$ matrix with the ith row given by $\{\nabla f_i(\boldsymbol{\theta})\}^T$ and $\mathbf{y}^*(\boldsymbol{\theta})$ is an $n \times 1$ vector with the ith component $y_i - f_i(\boldsymbol{\theta}) + \{\nabla f_i(\boldsymbol{\theta})\}^T \boldsymbol{\theta}$. Thus h_n may

be written as a sum of its argument and a remainder term as $h_n(\boldsymbol{\theta}) = \boldsymbol{\theta} + R_n(\boldsymbol{\theta})$. The estimation will be exact if $R_n = 0$.

The Gauss-Newton iteration step is then given by.

$$\widehat{\boldsymbol{\theta}}_{W,n}^{(k)} = h_n(\widehat{\boldsymbol{\theta}}_{W,n}^{(k-1)}) \,. \tag{3.28}$$

for $k = 1, 2, \ldots$, where $\boldsymbol{\theta}^{(0)}$ is an initial estimate. This k-step estimate has the same asymptotic property as $\widehat{\boldsymbol{\theta}}_{W,n}$ as shown in the following theorem.

Theorem 3.4.1. *If the initial estimate is such that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n}^{(0)} - \boldsymbol{\theta}_0)$ is bounded in probability, then*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n}^{(k)} - \widehat{\boldsymbol{\theta}}_{W,n}) \xrightarrow{\mathcal{P}} \mathbf{0}.$$

for any $k \geq 1$.

Proof. Applying Theorem 4.2 of McKean and Hettmansperger (1978), under the condition of the theorem, we have,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{W,n}^{(k)} - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{P}} \mathbf{0} \,.$$

for any $k \geq 1$. Lemma 3.3.1 completes the proof. □

Since $\tilde{\boldsymbol{\theta}}_n$ satisfies,

$$\mathbf{X}^{*T}\{G(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}_0) - 1/2\} = \mathbf{0} \,,$$

Lemma 3.3.1 and the representation in (3.25) give,

$$\mathbf{h}_n(\widehat{\boldsymbol{\theta}}_{W,n}) = \widehat{\boldsymbol{\theta}}_{W,n} + o_p(n^{-1/2}). \tag{3.29}$$

Thus $\widehat{\boldsymbol{\theta}}_{W,n}$ is a fixed point of the mapping $\mathbf{h}_n(\boldsymbol{\theta})$ when n is sufficiently large. As shown in Jennrich (1969), this is one of the two sufficient conditions for the asymptotic numerical stability of a Gauss-Newton type procedure. The second sufficient condition is that there exists a neighborhood $N_\delta(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ such that.

$$\left\| \frac{\partial \mathbf{h}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\| \leq c < 1 .$$

for $\boldsymbol{\theta} \in N_\delta(\boldsymbol{\theta}_0)$, when n is sufficiently large. Assuming the second derivatives $\nabla^2 f_i(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$ on Θ we have a $\delta > 0$ such that.

$$\left\| \frac{\partial \mathbf{h}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} - \frac{\partial \mathbf{h}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} \right\| \leq 1/4 .$$

for $\boldsymbol{\theta} \in N_\delta(\boldsymbol{\theta}_0)$ and large n. What remains to show is that $\|\partial \mathbf{h}_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^T\| \leq k < 3/4$ for large n. But we can show that $\|\partial \mathbf{h}_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^T\|$ goes to 0 in probability, which, of course, implies that $\|\partial \mathbf{h}_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^T\| \leq 1/2$ for sufficiently large n. Thus we have proven the following numerical stability theorem.

Theorem 3.4.2. *In addition to A1-A5 and N1-N3, assume that the second derivatives $\nabla^2 f_i(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}$ on Θ. Then there exist numbers $\delta > 0$ and n_0 such that the Gauss-Newton iteration given by* (3.28) *converges to $\widehat{\boldsymbol{\theta}}_{W,n}$ for any starting value in the spherical neighborhood $N_\delta(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ with radius δ.*

At each Gauss-Newton step we fit a linear regression model with no intercept parameter. This follows the projection technique of Dixon and McKean (1996). Suppose we need to estimate $\boldsymbol{\theta}^*$ of the model

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} , \tag{3.30}$$

but we fit the model

$$\mathbf{y}^* = \mathbf{1}\alpha_1 + \mathbf{X}^*\boldsymbol{\theta}_1^* + \boldsymbol{\varepsilon}\,. \tag{3.31}$$

where the true α_1 is 0. Let $\widehat{\mathbf{y}_1^*} = \mathbf{1}\widehat{\alpha_1} + \mathbf{X}^*\widehat{\boldsymbol{\theta}^*}_1$ be the Wilcoxon fitted value. Dixon and McKean (1996) show that the Wilcoxon fitted value based on (3.30) is the LS projection of $\widehat{\mathbf{y}_1}$ onto the space spanned by the columns of $\mathbf{X}^*$. Thus

$$\widehat{\boldsymbol{\theta}^*} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\widehat{\mathbf{y}_1^*}\,.$$

Therefore, every step of the Gauss-Newton iteration involves attaching a column of ones to the design matrix, fitting (3.31), and projecting onto the right space.

The stopping rule utilized looks at the difference between two consecutive estimates. For $\epsilon > 0$, $\|\widehat{\boldsymbol{\theta}}_{W,n}^{(k+1)} - \widehat{\boldsymbol{\theta}}_{W,n}^{(k)}\| < \epsilon$ implies that $\|R_n(\widehat{\boldsymbol{\theta}}_{W,n}^{(k)})\| < \epsilon$ by the continuity and asymptotic linearity (see Jurečková (1969)) of R_n. Since $\widehat{\boldsymbol{\theta}}_{W,n}$ is a fixed point for sufficiently large n, it follows that $\widehat{\boldsymbol{\theta}}_{W,n}^{(k)}$ is close to $\widehat{\boldsymbol{\theta}}_{W,n}$. Therefore we fit tangent linear models recursively until the convergence criterion is met. The linear models may be fit using the Robust General Linear Model (RGLM) package of Kapenga et. al. (1995). An alternative approach of obtaining R estimates of linear regression coefficients is using iteratively reweighted least square estimates as discussed in Sievers and Abebe (2002). The latter approach gives the investigator the flexibility of using any statistical package that fits linear models via LS to obtain R estimates of regression coefficients.

CHAPTER IV

GENERALIZED SIGNED-RANK ESTIMATION

4.1 Definition and Existence

Consider the following general regression model (model (1.2))

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \varepsilon_i, \qquad 1 \leq i \leq n\ , \tag{4.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is a vector of parameters, $\mathbf{x}_i \in \mathcal{X}$ is a vector of independent variables, and f is a real-valued function defined on $\boldsymbol{\Theta} \times \mathcal{X}$.

We shall assume that $\boldsymbol{\Theta}$ is compact, $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$, and $f(\mathbf{x}, \boldsymbol{\theta})$ is defined and continuous for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and is measurable for each $\mathbf{x} \in \mathcal{X}$.

We define the estimator of $\boldsymbol{\theta}_0$ to be any vector $\boldsymbol{\theta}$ minimizing

$$D_n^{\rho}(\mathbf{y}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} a_n(i)\rho(|z(\boldsymbol{\theta})|_{(i)}) \tag{4.2}$$

where $z_i(\boldsymbol{\theta}) = y_i - f(\mathbf{x}_i, \boldsymbol{\theta})$ and $|z(\boldsymbol{\theta})|_{(i)}$ is the ith ordered value among $|z_1(\boldsymbol{\theta})|, \ldots, |z_n(\boldsymbol{\theta})|$. The function $\rho : \Re^+ \to \Re^+$ is Borel measurable, continuous, and strictly increasing. The numbers $a_n(i)$ are scores generated as $a_n(i) = \varphi^+(i/(n+1))$, for some score function $\varphi^+ : (0, 1) \to \Re^+$.

This estimator will be denoted by $\widehat{\boldsymbol{\theta}}_{\rho,n}$.

The general dispersion function D_n^{ρ} contains infinitely many possible dispersion functions depending on the functions φ^+ and ρ examples of which are

the L_1 and LS dispersion functions. By truncating the ends of the score generating function, φ^+, we obtain high breakdown estimates which are very helpful when the data contain outliers. This chapter develops asymptotic and robustness properties of $\widehat{\boldsymbol{\theta}}_{\rho,n}$.

Theorem 4.1.1. *Under model* (4.1), $\widehat{\boldsymbol{\theta}}_{\rho,n}$ *exists.*

Proof. Because $D_n^\rho(\mathbf{y}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, Lemma 3.1.1 implies the existence of a minimizer of $D_n^\rho(\mathbf{y}, \boldsymbol{\theta})$. □

In this chapter we use the notation introduced in Section 2.4. Recall that we denote by L^p, $1 \leq p \leq \infty$, the space of measurable functions $h : (0,1) \to \Re$ for which $|h|^p$ is integrable for $1 \leq p < \infty$ and the space of essentially bounded measurable functions for $p = \infty$. The L^p norm of h is

$$\|h\|_p \equiv \begin{cases} \{\int |h|^p\}^{1/p} & \text{if } 1 \leq p < \infty \text{ , and} \\ \text{ess sup } |h| & \text{if } p = \infty \text{ .} \end{cases} \tag{4.3}$$

All integrals are with respect to Lebesgue measure on $(0,1)$. The range of integration will be assumed to be $(0,1)$ unless specified otherwise.

4.2 Strong Consistency

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $i = 1, \ldots, n$, assume that $\mathbf{x}_i$ and $\varepsilon_i = y_i - f(\mathbf{x}_i; \boldsymbol{\theta}_0)$ are independent random variables (carried by $(\Omega, \mathcal{F}, P)$) with

distributions H and G. respectively. We shall write $\mathbf{x}.\ e$ and $|z(\boldsymbol{\theta})|$ for $\mathbf{x}_1,\ \varepsilon_1$ and $|z_1(\boldsymbol{\theta})|$ respectively.

We will assume

S1: $P(f(\mathbf{x}:\boldsymbol{\theta}) = f(\mathbf{x}:\boldsymbol{\theta}_0)) < 1$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$;

S2: for $1 \leq p.q \leq \infty$ with $1/p + 1/q = 1$ we have $E[\rho(|z(\boldsymbol{\theta})|)]^q < \infty$ and $\|\varphi^+\|_p < \infty$: and.

S3: G has a density g with a unique mode at 0.

As shown in Corollary 3.2.1. Assumption S1 is the condition needed for $\boldsymbol{\theta}_0$ to be identified. The linear version of S1 was given by Hössjer (1994) as $P(|\boldsymbol{\theta}'\mathbf{x}| = 0) < 1$ under the assumption that $\boldsymbol{\theta}_0 = 0$. As we will see in Section 4.3. other works on nonlinear regression assume conditions which are stronger than S1.

If $z(\boldsymbol{\theta})$ follow a distribution $\tilde{G}_{\boldsymbol{\theta}}$. then S2 puts $\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1}$ and φ^+ in conjugate spaces when $p \in (1.\infty)$. Hölder's inequality ensures that the product $(\varphi^+)(\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1})$ is integrable. Furthermore. if ρ is a convex function. an application of Minkowski's inequality yields

$$\{E[\rho(|z(\boldsymbol{\theta})|)]^q\}^{1/q} \leq \{E[\rho(|\varepsilon|)]^q\}^{1/q} + \{E[\rho(|f(\mathbf{x}:\boldsymbol{\theta}) - f(\mathbf{x}:\boldsymbol{\theta}_0)|)]^q\}^{1/q}.$$

Thus separate conditions on ε and f may be sufficient for $E[\rho(|z(\boldsymbol{\theta})|)]^q < \infty$.

Condition S3 admits a wide variety of error distributions examples of which are the normal, double exponential and Cauchy distributions with location parameter equal to 0. This replaces the 0 median assumption of Chapter III which was one of the conditions needed for the consistency of the Wilcoxon estimator. There is also no symmetry assumption placed on the error distribution.

We will now give some preliminary results needed for showing the consistency of $\widehat{\boldsymbol{\theta}}_{\rho,n}$. For our purposes let $J_n(t) = \varphi^+(i/(n+1))I_{((i-1)/n, i/n]}(t)$ for $i = 1, \ldots, n$ where I_A is the indicator of the set A. Notice that J_n is a step function and thus the uniform integrabilty condition in assumption (ii) of Lemma 2.4.1 becomes

$$\lim_{\alpha \to \infty} \sup_n \frac{1}{n} \sum_{i \in A_\alpha} |\varphi^+(i/(n+1))| = 0 \, .$$

where $A_\alpha = \{j : |\varphi^+(j/(n+1))| > \alpha\}$. This condition is satisfied if we have convergence in L^1 of J_n (see Theorem VI.18 of Doob (1994)). To this end, we will marginally violate assumption (ii) of Lemma 2.4.1 and assume that

$$\sup_n \|J_n\|_p \equiv \sup_n \left\{ \frac{1}{n} \sum_{i=1}^{n} |\varphi^+(i/(n+1))|^p \right\}^{1/p} < \infty \, . \tag{4.4}$$

for $1 \leq p \leq \infty$. Notice also that

$$\frac{1}{n} \sum_{i=1}^{[nt]} \varphi^+(i/(n+1)) \leq \int_0^t J_n \leq \frac{1}{n} \sum_{i=1}^{[nt]+1} \varphi^+(i/(n+1)) \, ,$$

where $[b]$ stands for the greatest integer less than or equal to b. Taking the limit as $n \to \infty$ we obtain that $\lim_{n \to \infty} \int_0^t J_n = \int_0^t \varphi^+$ for all $t \in (0, 1)$ provided that φ^+ has at most a finite number of discontinuities. Thus if φ^+ satisfies (4.4) and

$g \in L^q$ all the conditions of Lemma 2.4.1 hold. The following corollary is a special case of this result.

Corollary 4.2.1. *Let $W_1, \ldots, W_n$ be a random sample from a distribution F with support on $\Re^+$. Let $\rho : \Re^+ \to \Re^+$ be a continuous Borel measurable function. Suppose, for $1 \leq p, q \leq \infty$ with $1/p + 1/q = 1$, $E[\rho(W)]^q < \infty$ and $\|\varphi^+\|_p < \infty$. Then*

$$T_n \equiv n^{-1} \sum_{i=1}^{n} \varphi^+(i/(n+1))\rho(W_i) \xrightarrow{a.s.} \int (\varphi^+)\ (\rho \circ F^{-1}) < \infty\ .$$

Proof. Let $g = \rho \circ F^{-1}$. Because $E[\rho(W)]^q < \infty$, a simple change of variable shows that $g \in L^q$. Following the arguments preceding the corollary, and applying Lemma 2.4.1, gives the desired result. □

Lemma 4.2.1. *Under assumptions S1 - S3*

$$D_n^\rho(\mathbf{y}, \boldsymbol{\theta}) \xrightarrow{a.s.} \mu(\boldsymbol{\theta}), \tag{4.5}$$

where $\mu : \Theta \to \Re$ is a function satisfying

$$\inf_{\boldsymbol{\theta} \in \Theta^*} \mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0) \quad a.s. \tag{4.6}$$

for any Θ^* a closed subset of Θ not containing $\boldsymbol{\theta}_0$.

Proof. The expression (4.5) follows from Corollary 4.2.1 which also furnishes the function

$$\mu(\boldsymbol{\theta}) \equiv \int (\varphi^+)\ (\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1}) < \infty$$

where $\tilde{G}_{\boldsymbol{\theta}}$ is the distribution function of $|z(\boldsymbol{\theta})|$.

To establish (4.6) we follow a similar strategy as in Hössjer (1994). Under S1 and S3 for any $s > 0$, for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

$$\begin{aligned}\tilde{G}_{\boldsymbol{\theta}}(s) &= P(|\varepsilon - \{f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)\}| \leq s) \\ &= E_{\mathbf{x}}\{P_{\varepsilon}(|\varepsilon - \{f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)\}| \leq s|\mathbf{x})\} \\ &< E_{\mathbf{x}}\{P_{\varepsilon}(|\varepsilon| \leq s)\} = \tilde{G}_{\boldsymbol{\theta}_0}(s)\end{aligned}$$

Since μ is a continuous function depending on $\boldsymbol{\theta}$ only through $\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1}$ and since ρ is a strictly increasing function, it follows that $\mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0)$ whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Thus for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$, we have a $\mu^* \in \Re$ such that $\mu(\boldsymbol{\theta}) > \mu^* > \mu(\boldsymbol{\theta}_0)$. Therefore it must be true that $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0)$ a.s. □

The following theorem gives the strong consistency of $\widehat{\boldsymbol{\theta}}_{\rho,n}$.

Theorem 4.2.1. *Under S1 - S3.* $\widehat{\boldsymbol{\theta}}_{\rho,n} \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

Proof. By Lemma 2.1.2, to establish the consistency of $\widehat{\boldsymbol{\theta}}_{\rho,n}$, it is sufficient to show that

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \left(D_n^{\rho}(\mathbf{y},\boldsymbol{\theta}) - D_n^{\rho}(\mathbf{y},\boldsymbol{\theta}_0)\right) > 0 \quad \text{a.s.}, \tag{4.7}$$

for any $\boldsymbol{\Theta}^*$ a closed subset of $\boldsymbol{\Theta}$ not containing $\boldsymbol{\theta}_0$. But

$$\begin{aligned}\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \left(D_n^{\rho}(\mathbf{y},\boldsymbol{\theta}) - D_n^{\rho}(\mathbf{y},\boldsymbol{\theta}_0)\right) \geq \liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} A_n(\boldsymbol{\theta}) + \\ \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} B(\boldsymbol{\theta},\boldsymbol{\theta}_0) + \liminf_{n\to\infty} C_n(\boldsymbol{\theta}_0)\,,\end{aligned} \tag{4.8}$$

where

$$A_n(\boldsymbol{\theta}) = D_n^{\rho}(\mathbf{y}, \boldsymbol{\theta}) - \mu(\boldsymbol{\theta}) ,$$

$$B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta}_0) , and$$

$$C_n(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}_0) - D_n^{\rho}(\mathbf{y}, \boldsymbol{\theta}_0) .$$

As a result of Corollary 4.2.1, $\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} A_n(\boldsymbol{\theta})$ and $\liminf_{n\to\infty} C_n(\boldsymbol{\theta}_0)$ are 0 a.s. Due to Lemma 4.2.1 we have $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > 0$ a.s. Therefore the statement given in (4.7) holds. The proof is complete. □

4.3 Some Corollaries

Next some special cases of interest are considered. We consider the L_1, least squares, signed-rank Wilcoxon, and their trimmed variations. We also look at a case where the score function is the inverse of a Gaussian distribution. All these cases involve a convex ρ and hence Minkowski's inequality may be used to supply assumption S2. Trimming is implemented by "chopping-off" the ends of the score generating function, φ^+ (see Hössjer (1994)). The proofs follow from Theorem 4.2.1 in a straightforward manner.

4.3.1 Least Squares, Least Trimmed Squares

The LS estimate of $\boldsymbol{\theta}_0$ is the value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that minimizes $\sum_{i=1}^{n} \{z_i(\boldsymbol{\theta})\}^2$. In this subsection we provide sufficient conditions for the strong consistency of the LS estimator via the objective function given in (4.2). We will also consider

the least trimmed sum of squares (LTS) estimator which minimizes the sum of the first $[\gamma n]$, $\gamma \in (1/2, 1)$, ordered squared residuals. The value of γ is usually taken such that $[\gamma n] = [n/2] + 1$ (see Rousseeuw (1983)) to provide estimators with high breakdown point.

Let $I_A(\omega)$ be a function such that $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. Let $\varphi^+(u) = I_{(\alpha,\beta)}(u)$ for $0 \leq \alpha < \beta \leq 1$ and $\rho(w) = w^2$ for $w \geq 0$. In the case where $\alpha = 0$ and $\beta = 1$ the dispersion function given by (4.2) is the least squares dispersion function. If there exist two positive real numbers, $0 < c_1 < c_2 < 1$, such that $c_1 < \alpha < \beta < c_2$, then the dispersion function becomes the least trimmed squares dispersion. The following corollary gives the sufficient conditions for the strong consistency of the least squares estimator.

Corollary 4.3.1. *If*

B1: $P(f(\mathbf{x}:\boldsymbol{\theta}) = f(\mathbf{x}:\boldsymbol{\theta}_0)) < 1$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

B2: $E(\varepsilon^2) < \infty$ *and* $E([f(\mathbf{x}:\boldsymbol{\theta}) - f(\mathbf{x}:\boldsymbol{\theta}_0)]^2) < \infty$ *for all* $\boldsymbol{\theta} \in \Theta$, *and*

B3: *G has a density g with a unique mode at 0,*

then the least squares (least trimmed squares) estimator is strongly consistent for $\boldsymbol{\theta}_0$.

Proof. Assumption B1 is equivalent to S1 and assumption B3 is equivalent to S3. We need to show that B2 implies S2. Let $q = 1$ and $p = \infty$ so L^q and L^p are

conjugate spaces. This implies that $E[\rho(|z(\boldsymbol{\theta})|)]^q = E[z^2(\boldsymbol{\theta})] < \infty$ by B2 and Minkowski's inequality. What remains to show is that $\|\varphi^+\|_\infty$ is bounded. But $\|\varphi^+\|_\infty = \text{ess sup } |\varphi^+| = \sup\{c : \{|\varphi^+| > c\} \neq \emptyset\}$. From the definition of φ^+ it follows that $\|\varphi^+\|_\infty = \sup\{c : \{I_{(\alpha,\beta)}(u) > c\} \neq \emptyset\} = 1$. □

Jennrich (1969) establishes the strong consistency of the least squares estimator under some assumptions. His assumptions in the notation of this paper are

J1: $E([f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)]^2) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

J2: $E(\varepsilon^2) < \infty$ and $E([f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)]^2) < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and

J3: $E(\varepsilon) = 0$.

Assumptions B2 and J2 are identical. B3 and J3, while not generally comparable, are identical in most practical situations where a symmetric, unimodal error density is assumed. As the next proposition shows, B1 is weaker than J1 in the sense that whenever J1 is true B1 is also true.

Proposition 4.3.1. *B1 is weaker than J1 in the sense that J1 $\Rightarrow$ B1.*

Proof. Assume that B1 fails to hold, that is there exists a point $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$ in $\boldsymbol{\Theta}$ such that $P(f(\mathbf{x};\boldsymbol{\theta}') = f(\mathbf{x};\boldsymbol{\theta}_0)) = 1$. This implies that $E([f(\mathbf{x};\boldsymbol{\theta}') - f(\mathbf{x};\boldsymbol{\theta}_0)]^2) = 0$. Thus J1 fails. Therefore, J1 implies B1. □

4.3.2 L_1, Trimmed Absolute Deviations

The L_1 estimator corresponds to the case where $\varphi^+ \equiv 1$ and $\rho(w) = w$ for $w \geq 0$. A situation similar to the least trimmed squares estimator holds for the trimmed absolute deviations estimator. The sufficient conditions for the strong consistency of the L_1 and trimmed absolute deviations estimators are given in the following corollary.

Corollary 4.3.2. *If*

C1: $P(f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}_0)) < 1$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

C2: $E(|\varepsilon|) < \infty$ *and* $E(|f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)|) < \infty$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *and*

C3: G *has a density* g *with a unique mode at 0,*

then the L_1 *(trimmed absolute deviations) estimator is strongly consistent for* $\boldsymbol{\theta}_0$.

Proof. The proof is similar to the proof of Corollary 4.3.1. □

We next compare the result in Corollary 4.3.2 with the one given by Oberhofer (1982). Although his conditions were sufficient to give the strong consistency of the L_1 estimator via Lemma 2.1.1, Oberhofer proves *weak* consistency by imposing the following conditions.

O1: If Θ^* is a closed set not containing θ_0, then there exist numbers $\epsilon > 0$ and n_0 such that for all $n \geq n_0$

$$\inf_{\theta \in \Theta^*} n^{-1} \sum_{i=1}^{n} |l_i(\theta)| \min\{G(|l_i(\theta)|/2) - 1/2\ ,\ 1/2 - G(-|l_i(\theta)|/2)\} \geq \epsilon\ ,$$

for all such Θ^* where $l_i(\theta) = f(\mathbf{x}_i; \theta) - f(\mathbf{x}_i; \theta_0)$.

O2: $E(|\varepsilon|) < \infty$ and $E([f(\mathbf{x}; \theta) - f(\mathbf{x}; \theta_0)]^2) < \infty$ for all $\theta \in \Theta$, and

O3: $G(0) = 1/2$.

Once again C3 and O3 are not comparable. O2 is stronger than C2. Following similar contrapositive arguments as in Proposition 4.3.1, we can easily show that O1 is also stronger than C1. For a detailed discussion of this and sufficient conditions for O1, the reader is referred to Oberhofer (1982). One can immediately observe that the identifiability of θ_0 is a neccesary condition for O1.

4.3.3 Signed-Rank

The signed-rank norm is given by

$$\frac{1}{n} \sum_{i=1}^{n} \varphi^+\left(\frac{i}{n+1}\right) |z(\theta)|_{(i)}\ .$$

This norm was considered by Hössjer (1994) to provide an estimator with a positive breakdown in the linear model. Here we give a nonlinear analogue of that result.

Set $\varphi^+(u) = u$ for $0 < u < 1$ and $\rho(w) = w$ for $w \geq 0$. The following corollary gives the sufficient conditions for the strong consistency of the signed-rank Wilcoxon estimator. The proof is analogous to the proof of Corollary 4.3.2.

Corollary 4.3.3. *If*

D1: $P(f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}_0)) < 1$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

D2: *for some* $r > 1$, $E(|\varepsilon|^r) < \infty$ *and* $E(|f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)|^r) < \infty$ *for all* $\boldsymbol{\theta} \in \Theta$, *and*

D3: *G has a density g with a unique mode at 0,*

then the signed-rank estimator is strongly consistent for $\boldsymbol{\theta}_0$.

Proof. D1 and D3 are equivalent to S1 and S3, respectively. To show that S2 is true whenever D2 is true, let $q = r > 1$. We need to show that $\|\varphi^+\|_p < \infty$ for any $p \in (1, \infty)$, which is obvious since φ^+ is the identity function. □

4.3.4 Normal Scores

The frequently used normal scores are generated by

$$\varphi^+(u) = \Phi^{-1}\Big(\frac{u+1}{2}\Big),$$

for $u \in (0, 1)$ where Φ represents the standard normal distribution function. We will refer to the value in Θ minimizing (4.2) with such φ^+ and $\rho(w) = w$ as the

normal scores estimator. These scores were first proposed by Fraser (1957). The following corollary shows that conditions D1 - D3 given above are sufficient for consistency of the normal scores estimator.

Corollary 4.3.4. *Under D1 - D3, the normal scores estimator is strongly consistent.*

Proof. In light of the proof of Corollary 4.3.3, we need only show that $\|\varphi^+\|_p < \infty$ for any $p \in (1, \infty)$ and $\varphi^+(u) = \Phi^{-1}((u+1)/2)$. Let $2k$ be the smallest even integer greater than or equal to p. Via a change of varable we can see that $\|\varphi^+\|_p^p = E[T^p]$ where T is a random variable that follows the standard normal distribution. Using the moment generating function one can easily show that (see Lehmann (1997)),

$$\|\varphi^+\|_p^p \leq E[T^{2k}] = \frac{(2k)!}{2^k k!} .$$

This quantity is finite for any $k < \infty$. □

4.4 Breakdown Point

One of the virtues of the estimators discussed in this paper is that they allow for trimming. This in turn provides us with estimates that are robust when one or more of the model assumptions are violated. In this section we will consider the breakdown point of our estimator as a measure of its robustness.

Let $Z = \{(x_1, y_1), ..., (x_n, y_n)\}$ denote the sample data points. Let $\mathcal{Z}^m$ be the set of all data sets obtained by replacing any m points in Z by arbitrary

points. The finite sample breakdown point of an estimator $\widehat{\boldsymbol{\theta}}$ is defined as (see Donoho and Huber (1983))

$$\mathrm{BD}_n^*(\widehat{\boldsymbol{\theta}}, Z) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Z^m \in \mathcal{Z}^m} |\widehat{\boldsymbol{\theta}}(Z^m) - \widehat{\boldsymbol{\theta}}(Z)| = \infty \right\}, \tag{4.9}$$

where $\widehat{\boldsymbol{\theta}}(Z)$ is the estimate obtained based on the sample Z. In nonlinear regression, however, this definition of the breakdown point fails since BD* is not invariant to nonlinear reparameterizations. For a discussion of this see Stromberg and Ruppert (1992). We will adopt the definition of breakdown point for nonlinear models given by Stromberg and Ruppert (1992). The definition proceeds by defining finite sample upper and lower breakdown points, BD_+ and BD_-, which depend on the regression model, f. For any $\mathbf{x} \in \mathcal{X}$, the upper and lower breakdown points are defined as

$$\mathrm{BD}_+(f, \widehat{\boldsymbol{\theta}}, Z, \mathbf{x}) = \begin{cases} \min_{0 \le m \le n} \left\{ \frac{m}{n} : \sup_{Z^m \in \mathcal{Z}^m} f(\mathbf{x}, \widehat{\boldsymbol{\theta}}(Z^m)) = \sup_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) \right\} \\ \qquad \text{if } \sup_{\boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) > f(\mathbf{x}, \widehat{\boldsymbol{\theta}}), \\ 1 \qquad \text{otherwise,} \end{cases} \tag{4.10}$$

and

$$\mathrm{BD}_{-}(f,\widehat{\boldsymbol{\theta}},Z,\mathbf{x}) = \begin{cases} \min_{0\le m\le n}\left\{\frac{m}{n} : \inf_{Z^m\in\mathcal{Z}^m} f(\mathbf{x},\widehat{\boldsymbol{\theta}}(Z^m)) = \inf_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta})\right\} \\ \qquad \text{if } \inf_{\boldsymbol{\theta}} f(\mathbf{x},\boldsymbol{\theta}) < f(\mathbf{x},\widehat{\boldsymbol{\theta}}), \\ 1 \qquad \text{otherwise .} \end{cases} \tag{4.11}$$

Let

$$\mathrm{BD}(f,\widehat{\boldsymbol{\theta}},Z,\mathbf{x}) = \min\{\mathrm{BD}_{+}(f,\widehat{\boldsymbol{\theta}},Z,\mathbf{x}),\mathrm{BD}_{-}(f,\widehat{\boldsymbol{\theta}},Z,\mathbf{x})\} .$$

The finite sample breakdown point is now defined as

$$\mathrm{BD}(f,\widehat{\boldsymbol{\theta}},Z) = \inf_{\mathbf{x}\in\mathcal{X}}\left\{\mathrm{BD}(f,\widehat{\boldsymbol{\theta}},Z,\mathbf{x})\right\} . \tag{4.12}$$

The finite sample upper and lower breakdown points are defined analogously by replacing BD by BD_{+} and BD_{-}, respectively, in the above definition. Stromberg and Ruppert (1992) also show that $\mathrm{BD} = \mathrm{BD}^{*}$ in the case of a linear regression (i.e. $f(\mathbf{x},\boldsymbol{\theta}) = \mathbf{x}'\boldsymbol{\theta}$) and $\mathrm{BD} = n^{-1}$ for nonlinear least squares regression as expected.

Assume the scores $a_n(i)$ are nonnegative and

$$k = \max\{i : a_n(i) > 0\} ,$$

where $k \ge [n/2] + 1$. This is equivalent to

$$\sup\{u : \varphi^{+}(u) > 0\} \in (1/2, 1] .$$

This forces at least the first half of the ordered absolute residuals to contribute to the dispersion function. In light of this, the dispersion function may be written

as

$$D_n^{\rho}(y,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{k} a_n(i)\rho(|z(\boldsymbol{\theta})|_{(i)}) .$$

The following theorem is a version of Theorem 3 of Stromberg and Ruppert (1992). We impose the same conditions but give the result in terms of k. The results given are for upper breakdown. Analogues for lower breakdown are straight forward. In the following, $\#(A)$ denotes the cardinality of the set A.

Theorem 4.4.1. *Assume for some fixed x*

(i) $\sup_{\boldsymbol{\theta}} f(x,\boldsymbol{\theta}) = \infty$, *and*

(ii) *there exist $\tau_k \subset \{i : 1 \le i \le n\}$ where $\#(\tau_k) = 2n - [n/2] - k$ such that*

$$\lim_{M\to\infty}\left\{\inf_{\{\boldsymbol{\theta}:f(x,\boldsymbol{\theta})>M\}}\left\{\inf_{i\in\tau_k} f(x_i,\boldsymbol{\theta})\right\}\right\} = \infty .$$

Then

$$BD_+(f,\widehat{\boldsymbol{\theta}}_{\rho,n},Z) \ge \frac{n-k+1}{n} .$$

Proof. Let $m = n - k$ and let $r_i(\boldsymbol{\theta})$ be the residuals when the data come from $\mathcal{Z}^m$, the space with m points contaminated and $z_i(\boldsymbol{\theta})$ be the residuals from the original data. Let $\widehat{\boldsymbol{\theta}}_{\rho,n}(\mathcal{Z}^m)$ be the minimizer of the objective function (4.2) when the data come from $\mathcal{Z}^m$. The set $\{\rho(|r(\boldsymbol{\theta})|_{(i)}) : i = 1,2,\cdots,k\}$ contains at least $k - m$ elements of $\{\rho(|z(\boldsymbol{\theta})|_{(i)}) : i = 1,\cdots,n\}$. But since $k - m + \#(\tau_k) \ge n + 1$, $\{\rho(|r(\boldsymbol{\theta})|_{(i)}) : i = 1,2,\cdots,k\}$ must also contain at least one element of $\{\rho(|z_i(\boldsymbol{\theta})|) :$

$i \in \tau_k\}$. Thus, by **(ii)** of the theorem,

$$\lim_{M\to\infty}\left\{\inf_{\{\boldsymbol{\theta}:f(x,\boldsymbol{\theta})>M\}}\left\{\sum_{i+1}^{k} a_n(i)\rho(|r(\boldsymbol{\theta})|_{(i)})\right\}\right\} = \infty\,.$$

This means that $f(x, \widehat{\boldsymbol{\theta}}_{\rho,n}(\mathcal{Z}^m))$ remains bounded below some finite M. Since $\sup_{\boldsymbol{\theta}} f(x,\boldsymbol{\theta}) = \infty$, perturbing m points does not cause upper breakdown. Thus,

$$\mathrm{BD}_+(f, \widehat{\boldsymbol{\theta}}_{\rho,n}, Z) \geq \frac{m+1}{n} = \frac{n-k+1}{n}\,.$$

□

The same expression can be obtained for lower breakdown by multiplying the quantities y and $f(x,\boldsymbol{\theta})$ by -1. Thus the finite sample breakdown point is at least equal to the proportion of residuals trimmed out of the dispersion function. It could be significantly higher depending on the choice of ρ. The lower bound is attained if the objective function is the LS objective function.

The following is an immediate corollary which gives the breakdown point of $\widehat{\boldsymbol{\theta}}_{\rho,n}$.

Corollary 4.4.1. *Let* $\alpha = \sup\{u : \varphi^+(u) > 0\}$ *be such that* $1/2 < \alpha \leq 1$. *The breakdown point of* $\widehat{\boldsymbol{\theta}}_{\rho,n}$ *is at least* $1-\alpha$.

Proof. In Theorem 4.4.1 above, take $k = [\alpha n]$. Then,

$$\lim_{n\to\infty}\frac{n-k+1}{n} = \lim_{n\to\infty}\frac{n-[\alpha n]+1}{n} = 1-\alpha\,.$$

□

CHAPTER V

WEIGHTED WILCOXON ESTIMATION

5.1 Definition and Existence

Consider the general regression model given in (1.1),

$$y_i = f_i(\boldsymbol{\theta}_0) + \varepsilon_i, \quad \text{for} \quad 1 \leq i \leq n\ ,$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^\circ$. Throughout this chapter we will assume that $\boldsymbol{\Theta}$ is a compact subspace of $\Re^p$.

We define the weighted Wilcoxon dispersion function by

$$D_n^w(\boldsymbol{\theta}) \equiv \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})|z_i(\boldsymbol{\theta}) - z_j(\boldsymbol{\theta})|\ , \tag{5.1}$$

where $z_i(\boldsymbol{\theta}) = y_i - f_i(\boldsymbol{\theta})$, $i = 1, \ldots, n$, $\widehat{\boldsymbol{\theta}}_n^{(0)}$ is an initial estimator, and w_{ij} are weight functions. We will denote the minimizer by $\widehat{\boldsymbol{\theta}}_{W,n}$.

The dispersion function given in (5.1) is a generalization of D_n considered in Chapter III. When all the weights are equal to unity, D_n^w reduces to D_n. Thus the results of Chapter III are special cases of the results of this chapter for the case where $w_{ij} \equiv 1$ for all $1 \leq i < j \leq n$.

Estimation based on the weighted Wilcoxon dispersion function was introduced by Sievers (1983) who assumed the weights, w_{ij}, to be non-stochastic.

Naranjo and Hettmansperger (1994) further developed the weighted Wilcoxon and used it to obtain the so-called generalized R (GR) estimates of regression coefficients in the linear model. By using Mallows weights they were able to obtain estimators with a bounded influence function.

The weighted Wilcoxon dispersion function (5.1) in its present form was given by Chang et. al. (1999). They considered the linear model, $f_i(\boldsymbol{\theta}_0) = \mathbf{x}_i^T\boldsymbol{\theta}_0$, and obtained estimates of $\boldsymbol{\theta}_0$ that have high breakdown point and at the same time possess high efficiency. They also show that if the initial estimator, $\widehat{\boldsymbol{\theta}}_n$, has high breakdown, then the estimate obtained by minimizing (5.1) will have high breakdown as well.

The following theorem gives the existence of the minimizer of (5.1).

Theorem 5.1.1. *Under model* (1.1), *if for* $1 \leq i < j \leq n$, $w_{ij}(\cdot)$ *are continuous,* $\Re^+$ *valued functions, then* $\widehat{\boldsymbol{\theta}}_{V,n}$ *exists.*

Proof. The proof follows immediately from Lemma 3.1.1 since D_n^w is a continuous nonnegative function. □

5.2 Consistency

The consistency of $\widehat{\boldsymbol{\theta}}_{V,n}$ will be shown under regularity conditions that are analogous to the ones used in Chapter III. Generally, these conditions reduce to the conditions of Chapter III when the weights are unity. The one difference is that A5 is replaced by a weighted version of its sufficient conditions. The intermediate

steps are incorporated in our proofs.

Once again consistency is established by appealing to Lemma 2.1.1 and Lemma 2.1.2, our general results. The notation of Chapter III will be assumed throughout this chapter.

The following assumptions will be needed.

WC1: For $1 \leq i, j \leq n$, w_{ij} are nonegative, continuous functions with $w_{ij}(\boldsymbol{\theta}_0) \leq M < \infty$, $\sum_{i,j} w_{ij}(\boldsymbol{\theta}_0) > 0$, and gradients ∇w_{ij} bounded uniformly in i and j on $\boldsymbol{\Theta}^\circ$.

WC2: $\lim_{n\to\infty} n^{-1}\Delta_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

WC3: $\widehat{\boldsymbol{\theta}}_n^{(0)} \rightarrow \boldsymbol{\theta}_0$ in probability.

The condition $\sum_{i,j} w_{ij}(\boldsymbol{\theta}_0) > 0$ in WC1 says that there is at least one w_{ij} that is strictly positive. WC1 also imposes a boundedness condition on the weight functions and the slope of the tangent lines at any point in $\boldsymbol{\Theta}^\circ$. Since the weight functions are $\Re^+$-valued, and continuous, they map a compact region to a closed and bounded subset of $\Re^+$. In practice the investigator needs to make sure that the weight functions are not capable of giving excessively large weights since the dispersion function will be inflated and the estimator will break down eventhough we may have a perfectly good data set.

The consistency of $\widehat{\boldsymbol{\theta}}_n^{(0)}$ given in WC3 is satisfied by taking any of the estimators discussed in Chapter III and Chapter IV. As mentioned above, since

the weighted Wilcoxon estimator borrows its robustness properties from $\widehat{\boldsymbol{\theta}}_n^{(0)}$, it will be wise to use any of the trimmed estimators given in Chapter IV rather than the Wilcoxon estimator which has an unbounded influence function and hence 0 breakdown. A common choice is the LTS estimator (see Hettmansperger and McKean (1998)).

In Lemma 3.2.1, it was shown that $D_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta}_0)$ minus its expectation converges (pointwise) to zero in probability. We anticipate a similar type of convergence for the process D_n^w given the weight functions and the initial estimator, $\widehat{\boldsymbol{\theta}}_n^{(0)}$, behave in a favorable manner. As the following lemma shows WC1 and WC3 are the conditions needed on w_{ij} and $\widehat{\boldsymbol{\theta}}_n^{(0)}$ to get the desired type of convergence.

Lemma 5.2.1. *Under assumptions WC1 - WC3,*

$$\{D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)\} - E\{D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)\} \rightarrow 0\,,$$

in probability.

Proof. Let $T_{ij} = |(\varepsilon_i - \varepsilon_j) + h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| - |\varepsilon_i - \varepsilon_j|$, $1 \leq i, j \leq n$. We may write

$$D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)}) T_{ij}$$

For an arbitrary $\delta > 0$, applying Markov's, Minkowski's and Jensen's inequalities, respectively, we have

$$
\begin{aligned}
&P\left(\left|\left[\binom{n}{2}\right]^{-1}\sum_{i<j}(w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij} - E(w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij})\right| > \delta\right) \\
&\qquad \le \frac{1}{\delta}\left[\binom{n}{2}\right]^{-1} E\left|\sum_{i<j}\{w_{ij}(\widehat{\boldsymbol{\theta}}_{W,n}^{(0)})T_{ij} - E(w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij})\}\right| \\
&\qquad \le \frac{1}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j} E|w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij} - E(w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij})| \\
&\qquad \le \frac{1}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j}(E|w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij}| + |E(w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij})|) \\
&\qquad \le \frac{2}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j} E|w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij}| \, .
\end{aligned}
$$

Now consider $w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})$. Expanding it about $\boldsymbol{\theta}_0$ using a Taylor Series approximation we get

$$
w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)}) = w_{ij}(\boldsymbol{\theta}_0) + [\nabla w_{ij}(\boldsymbol{\phi})]^T(\widehat{\boldsymbol{\theta}}_n^{(0)} - \boldsymbol{\theta}_0) \, , \tag{5.2}
$$

where $\boldsymbol{\phi} \in \boldsymbol{\Theta}^\circ$, with $\|\boldsymbol{\phi} - \boldsymbol{\theta}_0\| \le \|\widehat{\boldsymbol{\theta}}_n^{(0)} - \boldsymbol{\theta}_0\|$. Thus applying WC1, we have

$$
\begin{aligned}
&\frac{2}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j} E|w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})T_{ij}| \\
&\qquad \le \frac{2M}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j} E|T_{ij}| \\
&\qquad\quad + \frac{2}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j} E(|[\nabla w_{ij}(\boldsymbol{\phi})]^T(\widehat{\boldsymbol{\theta}}_n^{(0)} - \boldsymbol{\theta}_0)| \cdot |T_{ij}|) \\
&\qquad = I_{1n} + I_{2n} \, , \quad \text{say.}
\end{aligned}
$$

Consider I_{1n}.

$$
\begin{aligned}
I_{1n} &= \frac{2M}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j}E|T_{ij}| \\
&\le \frac{2M}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \\
&\le \frac{2M}{\delta}\left[\binom{n}{2}\right]^{-1}\left\{\sum_{i<j}|h_i^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)|+\sum_{i<j}|h_j^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\right\} \\
&= \frac{4M}{\delta n}\sum_{i=1}^{n}|h_i^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \\
&\le \frac{4M}{\delta}\left\{\frac{1}{n^2}\sum_{i=1}^{n}[h_i^*(\boldsymbol{\theta},\boldsymbol{\theta}_0)]^2\right\}^{1/2}.
\end{aligned}
$$

Hence, by WC2, I_{1n} goes to 0.

Considering I_{2n}, since $[\nabla w_{ij}(\boldsymbol{\phi})]^T(\widehat{\boldsymbol{\theta}}_n^{(0)}-\boldsymbol{\theta}_0)$ is $o_p(1)$, to prove $I_{2n}\to 0$ as $n\to\infty$, it suffices to show that

$$\frac{2}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j}E|T_{ij}|$$

is bounded.

But since $|T_{ij}|\le|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|$, following the arguments above, we have

$$\frac{2}{\delta}\left[\binom{n}{2}\right]^{-1}\sum_{i<j}E|T_{ij}|\to 0\,,$$

as $n\to\infty$. The proof is complete. □

In addition to WC1 - WC3 above, assume the following.

WC4: $\varepsilon_i-\varepsilon_j$ have a common distribution G which satisfies $G(0)=1/2$ and has density g continuous at 0 with $g(0)>0$.

WC5: For $1 \leq i, j \leq n$, and Θ^* a closed subset of $\Theta \setminus \{\boldsymbol{\theta}_0\}$, there exist a $\eta > 0$ and a n_0 such that for all $n \geq n_0$ we have

$$\inf_{\boldsymbol{\theta} \in \Theta^*} \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)(h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0))^2 \geq \eta \, .$$

Assumption WC4 is a combination of assumptions A4 and A5.1 in Chapter III. WC5 is a weighted version of A5.2. An alternative is to assume $w_{ij}(\boldsymbol{\theta}_0)$ are all bounded above zero in addition to A5.2. Even though most practical problems dictate that $w_{ij}(\boldsymbol{\theta}_0) > 0$ for all i, j, this imposes a much stronger assumption than WC1 and WC5 combined and hence will not be used here.

The lemma below is helpful when proving the convergence of a stochastic function via an application of a first order Taylor expansion.

Lemma 5.2.2. *Let T be an absolutely bounded random variable and $\mathbf{v} \in \Re^m$ be a non-stochastic vector with its components satisfying $|v_i| \leq v^* < \infty$ for $i = 1, \cdots, m$. Let $w_n = w_0 + \mathbf{v}^T \mathbf{z}_n$, where w_0 is a positive constant and $\mathbf{z}_n$ is a $\Re^m$-valued random variable with $\mathbf{z}_n \xrightarrow{p} \mathbf{0}$. Suppose there exists a n_0 such that $\mathbf{z}_n$ satisfies $|z_{ni}| \leq z^* < \infty$ for $i = 1, \cdots, m$ whenever $n \geq n_0$. Then*

$$E[w_n T] = w_0 E[T] + o(1) \, .$$

Proof. We need to show that $E[\mathbf{v}^T \mathbf{z}_n T] \to 0$. Since $|T| \leq T^* < \infty$ and $|v_i| \leq v^*$ we have

$$E[\mathbf{v}^T \mathbf{z}_n T] \leq |E[\mathbf{v}^T \mathbf{z}_n T]| \leq T^* v^* \sum_{i=1}^{m} E[|z_{ni}|] \, .$$

Let $\epsilon > 0$. Following Williams (1991), since $|z_{ni}| \xrightarrow{\mathcal{P}} 0$ for $i = 1, \cdots, m$, we may choose n_1 such that

$$P(|z_{ni}| > \epsilon/2) < \frac{\epsilon}{2z^*} .$$

whenever $n \geq n_1$.

Then, for $n \geq \max(n_0, n_1)$ and all i.

$$\begin{aligned} E[|z_{ni}|] &= E[|z_{ni}|I(|z_{ni}| > \epsilon/2)] + E[|z_{ni}|I(|z_{ni}| \leq \epsilon/2)] \\ &\leq z^* P(|z_{ni}| > \epsilon/2) + \epsilon/2 \\ &\leq \epsilon . \end{aligned}$$

The proof is complete. □

The following is a weighted version of Lemma 3.2.2.

Lemma 5.2.3. *Under WC4 and WC5, there exists a $\xi > 0$ and a n_0 such that for all $n \geq n_0$,*

$$\inf_{\boldsymbol{\theta} \in \Theta^*} E(D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)) \geq \xi .$$

Proof. Let $T_{ij} = [|(\varepsilon_i - \varepsilon_j) + h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| - |\varepsilon_i - \varepsilon_j|]$. One can easily observe that $|T_{ij}| \leq |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| < \infty$, for all $1 \leq i < j \leq n$. Thus, since

$$D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)}) T_{ij} ,$$

by the Lemma 5.2.2 and (3.10), we have

$$E(D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)) = 2\left[\binom{n}{2}\right]^{-1} \sum_{(i,j)\in A} w_{ij}(\boldsymbol{\theta}_0) \int_0^{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)} (|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| - x)dG(x)$$
$$+ 2\left[\binom{n}{2}\right]^{-1} \sum_{(i,j)\in B} w_{ij}(\boldsymbol{\theta}_0) \int_{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}^{0} (|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| + x)dG(x)$$
$$+ o(1) .$$

where A and B are a partition of $\{(i,j) : i < j\}$ according to $A = \{(i,j) : i < j \text{ and } h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) \leq 0\}$ and $B = \{(i,j) : i < j \text{ and } h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0) > 0\}$.

Restricting the ranges of integration and applying WC4 we get

$$E(D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)) \geq \left[\binom{n}{2}\right]^{-1} \sum_{(i,j)\in A} w_{ij}(\boldsymbol{\theta}_0)|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\left\{G\left(\frac{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}{2}\right) - \frac{1}{2}\right\}$$
$$+ \left[\binom{n}{2}\right]^{-1} \sum_{(i,j)\in B} w_{ij}(\boldsymbol{\theta}_0)|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\left\{\frac{1}{2} - G\left(\frac{-h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}{2}\right)\right\}$$
$$+ o(1)$$
$$\geq \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)|(h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \times$$
$$\min\left\{G\left(\frac{|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|}{2}\right) - \frac{1}{2}, \frac{1}{2} - G\left(\frac{-|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|}{2}\right)\right\} + o(1) .$$

Taylor expanding G about 0 and applying WC4 we get

$$E(D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)) \geq \frac{1}{2}\left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)(h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0))^2 g(t) ,$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$ and $t \in (-|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|/2, |h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|/2)$. Now since g is continuous at 0 and $g(0) > 0$, there is a symmetric neighborhood $(-\zeta,\zeta)$, $\zeta > 0$, over which $g^* \equiv \inf\{g(t) : t \in (-\zeta,\zeta)\} > 0$. Since $\boldsymbol{\Theta}^*$ is arbitrary, its boundary may be

chosen as close to $\boldsymbol{\theta}_0$ as desired. Thus, using WC5,

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} E(D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)) \geq \frac{\eta g^*}{2} = \xi > 0 .$$

The proof is complete. □

The following theorem gives the consistency of $\widehat{\boldsymbol{\theta}}_{V,n}$. The proof is similar to the proof of Theorem 3.2.1.

Theorem 5.2.1. *Under WC1 - WC5, $\widehat{\boldsymbol{\theta}}_{V,n}$ is weakly consistent for $\boldsymbol{\theta}_0$.*

Proof. The proof follows from Lemma 5.2.1 and Lemma 5.2.3 employing the same steps as the proof of Theorem 3.2.1. □

5.3 Asymptotic Normality

In this section we obtain the distributional properties of $\widehat{\boldsymbol{\theta}}_{V,n}$. This will follow a strategy similar to the one employed Chapter III.

We will start by defining

$$e_i^*(\boldsymbol{\theta}) = y_i - f_i(\boldsymbol{\theta}_0) + \{\nabla f_i(\boldsymbol{\theta}_0)\}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \quad for 1 \leq i \leq n , \tag{5.3}$$

where ∇f_i are the gradients as defined in Chapter III.

Let

$$y_i^* = (\mathbf{x}_i^*)^T\boldsymbol{\theta}_0 + \varepsilon_i ,$$

be as defined in (3.12). The quantities represented by $e_i^*(\boldsymbol{\theta})$ reduce to ε_i when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Define the associated weighted Wilcoxon dispersion function with deterministic weights as

$$T_n(\boldsymbol{\theta}) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)|e_i^*(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})| \,. \tag{5.4}$$

Denote the minimizer of T_n by $\widetilde{\boldsymbol{\theta}}_n$.

Notice that this is an estimator of a vector of linear regression coefficients. T_n corresponds to the dispersion function given by Sievers (1983). To prove the asymptotic normality of $\widehat{\boldsymbol{\theta}}_{V,n}$, we first show the symptotic equivalence of $\widehat{\boldsymbol{\theta}}_{V,n}$ and $\widetilde{\boldsymbol{\theta}}_n$ and then show the asymptotic normality of $\widetilde{\boldsymbol{\theta}}_n$.

Assume the following.

WN1: The true errors, ε_i, are independent, identically distributed with

$$E[|\varepsilon_1|] < \infty \,.$$

WN2: For $1 \leq i \leq n$ and $1 \leq j \leq p$, ∇f_{ij} are continuous in $\boldsymbol{\theta}$ on Θ°.

The lemma given below shows that under these assumptions, in addition to WC1-WC3, the estimators $\widehat{\boldsymbol{\theta}}_{V,n}$ and $\widetilde{\boldsymbol{\theta}}_n$ are asymptotically equivalent.

Lemma 5.3.1. *Under WC1-WC5, WN1, and WN2*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{V,n} - \widetilde{\boldsymbol{\theta}}_n) \;\rightarrow\; 0 \,,$$

in probability.

Proof. Consider

$$|D_n^w(\boldsymbol{\theta})-T_n(\boldsymbol{\theta})| = \left|\left[\binom{n}{2}\right]^{-1}\sum_{i<j}\left\{w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})|z_i(\boldsymbol{\theta})-z_j(\boldsymbol{\theta})|-w_{ij}(\boldsymbol{\theta}_0)|e_i^*(\boldsymbol{\theta})-e_j^*(\boldsymbol{\theta})|\right\}\right| .$$

For some $\boldsymbol{\phi} \in \Theta^\circ$, with $\|\boldsymbol{\phi}-\boldsymbol{\theta}_0\| \le \|\widehat{\boldsymbol{\theta}}_n^{(0)}-\boldsymbol{\theta}_0\|$, by (5.2) we have

$$\begin{aligned}
|D_n^w(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta})| &= \left[\binom{n}{2}\right]^{-1}\Bigg|\sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)\Big\{|z_i(\boldsymbol{\theta}) - z_j(\boldsymbol{\theta})| - |e_i^*(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})|\Big\} \\
&\qquad + \sum_{i<j}[\nabla w_{ij}(\boldsymbol{\phi})]^T(\widehat{\boldsymbol{\theta}}_n^{(0)} - \boldsymbol{\theta}_0)|z_i(\boldsymbol{\theta}) - z_j(\boldsymbol{\theta})|\Bigg| \\
&\le \left[\binom{n}{2}\right]^{-1}\sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)\Big|\{z_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta})\} - \{z_j(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})\}\Big| \\
&\qquad + \left[\binom{n}{2}\right]^{-1}\sum_{i<j}\Big|[\nabla w_{ij}(\boldsymbol{\phi})]^T(\widehat{\boldsymbol{\theta}}_n^{(0)} - \boldsymbol{\theta}_0)\Big| \cdot |z_i(\boldsymbol{\theta}) - z_j(\boldsymbol{\theta})| \\
&= C_{1n}(\boldsymbol{\theta}) + C_{2n}(\boldsymbol{\theta}) \text{ , say.} \qquad (5.5)
\end{aligned}$$

Consider $C_{1n}(\boldsymbol{\theta})$. By WC1 and the triangular inequality we have

$$\begin{aligned}
C_{1n}(\boldsymbol{\theta}) &\le M\left[\binom{n}{2}\right]^{-1}\left\{\sum_{i<j}|z_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta})| + \sum_{i<j}|z_j(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})|\right\} \\
&= \frac{2M}{n}\sum_{i=1}^{n}|z_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta})| .
\end{aligned}$$

By the definition of e_i^* we have

$$z_i(\boldsymbol{\theta}) - e_i^*(\boldsymbol{\theta}) = [\nabla f_i(\boldsymbol{\theta}_0)]^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) .$$

This implies that

$$\begin{aligned}
C_{1n}(\boldsymbol{\theta}) &\le \frac{2M}{n}\sum_{i=1}^{n}|[\nabla f_i(\boldsymbol{\theta}_0)]^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)| \\
&\le \frac{2M}{n}\max_{\substack{1\le i\le n\\ 1\le j\le p}}|\nabla f_{ij}(\boldsymbol{\theta}_0)|\sum_{j=1}^{p}|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j}| .
\end{aligned}$$

Since by WN2, ∇f_{ij} are $\Re$-valued, continuous functions defined on a compact space we have a number $K < \infty$ such that $|\nabla f_{ij}| \leq K$. Now taking the supremum of $C_{1n}(\boldsymbol{\theta})$ over a shrinking ball centered at $\widehat{\boldsymbol{\theta}}_{V,n}$ and radius $\delta/\sqrt{n}$ we get

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\leq\frac{\delta}{\sqrt{n}}} C_{1n}(\boldsymbol{\theta}) \leq \frac{2MK}{n}\left\{ \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\leq\frac{\delta}{\sqrt{n}}} \|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\| + \sum_{j=1}^{p} |(\widehat{\boldsymbol{\theta}}_{V,n})_j - \boldsymbol{\theta}_{0j}| \right\}$$
$$\leq \frac{2MK}{n}\left\{ \frac{\delta}{\sqrt{n}} + \sum_{j=1}^{p} |(\widehat{\boldsymbol{\theta}}_{V,n})_j - \boldsymbol{\theta}_{0j}| \right\}$$

Taking the limit as $n \to \infty$ and applying WC3, we have

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\leq\frac{\delta}{\sqrt{n}}} C_{1n}(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} 0\,. \tag{5.6}$$

Now consider $C_{2n}(\boldsymbol{\theta})$.

$$C_{2n}(\boldsymbol{\theta}) \leq \left[\binom{n}{2}\right]^{-1} \max_{\substack{1\leq i<j\leq n \\ 1\leq k\leq p}} \nabla w_{ijk}(\boldsymbol{\phi}) \sum_{k=1}^{p} |(\widehat{\boldsymbol{\theta}}_n^{(0)})_k - \boldsymbol{\theta}_{0k}| \sum_{i<j} |z_i(\boldsymbol{\theta}) - z_j(\boldsymbol{\theta})|$$
$$= \max_{\substack{1\leq i<j\leq n \\ 1\leq k\leq p}} \nabla w_{ijk}(\boldsymbol{\phi}) \left\{ \sum_{k=1}^{p} |(\widehat{\boldsymbol{\theta}}_n^{(0)})_k - \boldsymbol{\theta}_{0k}| \right\} \times$$
$$\left\{ \left[\binom{n}{2}\right]^{-1} \sum_{i<j} |(\varepsilon_i - \varepsilon_j) + h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \right\}$$
$$\leq K^* \left\{ \sum_{k=1}^{p} |(\widehat{\boldsymbol{\theta}}_n^{(0)})_k - \boldsymbol{\theta}_{0k}| \right\} \left\{ \frac{2}{n} \sum_{i=1}^{n} |\varepsilon_i| + \left[\binom{n}{2}\right]^{-1} \sum_{i<j} |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \right\},$$

where by WC1, $K^* < \infty$. Thus, since by WC3 $\widehat{\boldsymbol{\theta}}_n^{(0)} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$, to show that

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\leq\frac{\delta}{\sqrt{n}}} C_{2n}(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} 0\,,$$

we need only show that the quantity

$$\frac{2}{n} \sum_{i=1}^{n} |\varepsilon_i| + \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\leq\frac{\delta}{\sqrt{n}}} \left\{ \left[\binom{n}{2}\right]^{-1} \sum_{i<j} |h_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \right\},$$

remains bounded as $n \to \infty$. Using WN1 and applying the Weak Law of Large Numbers we have,

$$\frac{2}{n}\sum_{i=1}^{n}|\varepsilon_i| \xrightarrow{a.s.} 2E[|\varepsilon_1|] < \infty . \tag{5.7}$$

For $1 \le i \le n$, let

$$L_i = 2\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|f_i(\boldsymbol{\theta})| ,$$

and $L_{ij}^* = \max(L_i, L_j)$ for $1 \le i, j \le n$. Using the triangular inequality, for $1 \le i < j \le n$, we have

$$\begin{aligned}
\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| &\le \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|f_i(\boldsymbol{\theta})-f_i(\boldsymbol{\theta}_0)| + \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|f_j(\boldsymbol{\theta})-f_j(\boldsymbol{\theta}_0)| \\
&\le \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|f_i(\boldsymbol{\theta})-f_i(\widehat{\boldsymbol{\theta}}_{V,n})| + |f_i(\widehat{\boldsymbol{\theta}}_{V,n})-f_i(\boldsymbol{\theta}_0)| \\
&\quad + \sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|f_j(\boldsymbol{\theta})-f_j(\widehat{\boldsymbol{\theta}}_{V,n})| + |f_j(\widehat{\boldsymbol{\theta}}_{V,n})-f_j(\boldsymbol{\theta}_0)| . \\
&\le L_{ij}^*\left\{\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\| + \sum_{k=1}^{p}|(\widehat{\boldsymbol{\theta}}_{V,n})_k-\boldsymbol{\theta}_{0k}|\right\} \\
&\le L_{ij}^*\left\{\frac{\delta}{\sqrt{n}} + \sum_{k=1}^{p}|(\widehat{\boldsymbol{\theta}}_{V,n})_k-\boldsymbol{\theta}_{0k}|\right\} .
\end{aligned}$$

Since $L_{ij}^* < \infty$, an application of WC3 gives

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \xrightarrow{\mathcal{P}} 0 ,$$

which implies that

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}}\left\{\left[\binom{n}{2}\right]^{-1}\sum_{i<j}|h_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|\right\} \xrightarrow{\mathcal{P}} 0 . \tag{5.8}$$

Putting (5.7) and (5.8) together, we get

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V,n}\|\le\frac{\delta}{\sqrt{n}}} C_{2n}(\boldsymbol{\theta}) \xrightarrow{\mathcal{P}} 0 . \tag{5.9}$$

Expressions (5.6), (5.9), and (5.5) give

$$\sup_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|\leq\frac{\delta}{\sqrt{n}}} |D_n^w(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0 . \tag{5.10}$$

Now consider the distance $D_n^w(\boldsymbol{\theta}) - D_n^w(\widehat{\boldsymbol{\theta}}_{V:n}))$. We have

$$\inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} [D_n^w(\boldsymbol{\theta}) - D_n^w(\widehat{\boldsymbol{\theta}}_{V:n})] \geq [D_n^w(\boldsymbol{\theta}_0) - D_n^w(\widehat{\boldsymbol{\theta}}_{V:n})]$$

$$+ \inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} [D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)]$$

Because $\widehat{\boldsymbol{\theta}}_{V:n}$ is the minimizer of the continuous dispersion function $D_n^w(\boldsymbol{\theta})$, we get

$$[D_n^w(\boldsymbol{\theta}_0) - D_n^w(\widehat{\boldsymbol{\theta}}_{V:n})] \geq 0 .$$

An application of Lemma 5.2.1 and Lemma 5.2.3 yields

$$\inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} [D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)] \geq \inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} E[D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)]$$

$$+ \inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} \{D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0) - E[D_n^w(\boldsymbol{\theta}) - D_n^w(\boldsymbol{\theta}_0)]\}$$

$$\geq \xi > 0 .$$

whenever n is sufficiently large. This implies that

$$\inf_{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{V:n}\|=\frac{\delta}{\sqrt{n}}} [D_n^w(\boldsymbol{\theta}) - D_n^w(\widehat{\boldsymbol{\theta}}_{V:n})] \geq \xi > 0 . \tag{5.11}$$

Expressions (5.10) and (5.11) via an application of Lemma 2.2.1 give the desired result. □

Let $\gamma = \int h_\varepsilon^2$ where h_ε is the density of ε_1. Define

$$A_{ik}(\boldsymbol{\theta}_0) = \sum_{j=1}^{n} w_{ij}(\boldsymbol{\theta}_0)(\nabla f_{ij}(\boldsymbol{\theta}_0) - \nabla f_{jk}(\boldsymbol{\theta}_0)), \quad 1 \leq k \leq p,\ 1 \leq i \leq n .$$

Let A_n be the $n \times p$ matrix with the (i,k)th element equal to A_{ik} and let $V_n = A_n^T A_n$. Let $F_c = (I_n - n^{-1}J_n)\nabla f$ be the centered $n \times p$ design matrix and let $\overline{\nabla f}_k$ be the average of the kth column of ∇f. Here I_n is the $n \times n$ identity matrix while J_n is the $n \times n$ matrix of ones.

The following assumptions are given by Sievers (1983).

SN1: For $1 \leq i, j \leq n$, $w_{ij}(\cdot)$ are symmetric.

SN2: For each $k = 1, \ldots, p$,

$$\frac{\sum_{i=1}^{n} A_{ik}^2(\boldsymbol{\theta}_0)}{\max_{1 \leq i \leq n} A_{ik}^2(\boldsymbol{\theta}_0)} \rightarrow \infty.$$

SN3: For each $k = 1, \ldots, p$,

$$\frac{\sum_{i<j}[w_{ij}(\boldsymbol{\theta}_0)(\nabla f_{jk}(\boldsymbol{\theta}_0) - \nabla f_{ik}(\boldsymbol{\theta}_0))]^2}{\sum_{i=1}^{n} A_{ik}^2(\boldsymbol{\theta}_0)} \rightarrow 0.$$

SN4: For each $k = 1, \ldots, p$,

$$n^{-1/2} \max_{1 \leq i \leq n} |\nabla f_{ik}(\boldsymbol{\theta}_0) - \overline{\nabla f}_k(\boldsymbol{\theta}_0)| \rightarrow 0.$$

SN5: There is a positive definite matrix $\Sigma(\boldsymbol{\theta}_0)$ such that

$$n^{-1} F_c(\boldsymbol{\theta}_0)^T F_c(\boldsymbol{\theta}_0) \rightarrow \Sigma(\boldsymbol{\theta}_0).$$

SN6: There is a positive definite matrix $V(\boldsymbol{\theta}_0)$ such that

$$n^{-3} V_n(\boldsymbol{\theta}_0) \rightarrow V(\boldsymbol{\theta}_0).$$

SN7: For $k = 1, \ldots, p$, $2(n(n-1))^{-1} \sum_{i<j} [w_{ij}(\boldsymbol{\theta}_0)(\nabla f_{jk}(\boldsymbol{\theta}_0) - \nabla f_{ik}(\boldsymbol{\theta}_0))]^2$ is bounded as $n \to \infty$.

SN8: Let the $p \times p$ matrix $C_n(\boldsymbol{\theta}_0)$ be defined with the (k,l)th element

$$\sum_{i<j} w_{ij}(\boldsymbol{\theta}_0)(\nabla f_{jk}(\boldsymbol{\theta}_0) - \nabla f_{ik}(\boldsymbol{\theta}_0))(\nabla f_{jl}(\boldsymbol{\theta}_0) - \nabla f_{il}(\boldsymbol{\theta}_0)).$$

There is a nonsingular matrix C with $n^{-2}C_n(\boldsymbol{\theta}_0) \to C(\boldsymbol{\theta}_0)$.

The following theorem along with a proof can be found in Sievers (1983).

Theorem 5.3.1. *Under* (3.12), *WC4, SN1-SN8.*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (\gamma^2/12)C^{-1}(\boldsymbol{\theta}_0)V(\boldsymbol{\theta}_0)C^{-1}(\boldsymbol{\theta}_0)) .$$

We now give the main result of this section. Its proof is a direct application of Slutsky's Theorem.

Theorem 5.3.2. *Under model* (1.1), *WC1-WC5, WN1, WN2, SN1-SN8.*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{V,n} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (\gamma^2/12)C^{-1}(\boldsymbol{\theta}_0)V(\boldsymbol{\theta}_0)C^{-1}(\boldsymbol{\theta}_0)) .$$

Proof. The proof is an immediate consequence of Lemma 5.3.1 and Theorem 5.3.1.

□

Before concluding the present section, we present a proof of the asymptotic normality of $\hat{\boldsymbol{\theta}}_{V,n}$ which does not require the existence of $E(|\varepsilon_1|)$, an assumption which will not hold for some common probability distributions like the Cauchy distribution. This relaxation comes at a cost of making stronger assumptions

about the initial estimator, $\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}$, than just convergence in probability. The approach we follow starts out by showing the asymptotic equivalence of $\widehat{\boldsymbol{\theta}}_{V,n}$ and the estimator suggested by Chang et. al. (1999). Once the equivalence is established, the assumptions needed will be exactly those given by Chang et. al. (1999) in our notation.

Define

$$S_n(\boldsymbol{\theta}) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}) |e_i^*(\boldsymbol{\theta}) - e_j^*(\boldsymbol{\theta})| \,. \tag{5.12}$$

where e_i^* are as given in (5.3).

Let $\widehat{\widehat{\boldsymbol{\theta}}}_n$ denote a point in $\boldsymbol{\Theta}$ that minimizes S_n. This estimator is the one considered by Chang et. al. (1999) as a high breakdown estimator of linear regression coefficients. The following lemma shows that under the same regularity conditions considered in Lemma 5.3.1, except the existence of $E(|\varepsilon_1|)$, $\widehat{\widehat{\boldsymbol{\theta}}}_n$ and $\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}$ are asymptotically equivalent.

Lemma 5.3.2. *Under WC1-WC5 and WN2*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{V,n} - \widehat{\widehat{\boldsymbol{\theta}}}_n) \;\to\; 0 \,,$$

in probability.

Proof. By Lemma 2.2.1 and the inequality given in (5.11), we only need to show, for $\delta > 0$,

$$\sup_{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{V,n}\| \le \frac{\delta}{\sqrt{n}}} |D_n^w(\boldsymbol{\theta}) - S_n(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0 \,.$$

Let

$$K = \max_{\substack{1 \le i \le n \\ 1 \le j \le p}} |\nabla f_{ij}(\boldsymbol{\theta}_0)| .$$

By WN2, $K < \infty$. Following the combinatorial approach given in the proof of Lemma 5.3.1, one can show that

$$|D_n^w(\boldsymbol{\theta}) - S_n(\boldsymbol{\theta})| \le \frac{2KM}{n} \sum_{i=1}^{p} |\boldsymbol{\theta}_j - \boldsymbol{\theta}_{0j}| .$$

where $M < \infty$ is given in WC1. After an application of the triangular and Cauchy-Schwarz inequalities we get

$$\sup_{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{V,n}\| \le \frac{\delta}{\sqrt{n}}} |D_n^w(\boldsymbol{\theta}) - S_n(\boldsymbol{\theta})| \le \frac{2KM}{n} \left\{ \frac{\delta}{\sqrt{n}} + \sum_{j=1}^{p} |(\widehat{\boldsymbol{\theta}}_{V,n})_j - \boldsymbol{\theta}_{0j}| \right\} .$$

Because $\widehat{\boldsymbol{\theta}}_{V,n}$ is weakly consistent for $\boldsymbol{\theta}_0$, this implies

$$\sup_{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{V,n}\| \le \frac{\delta}{\sqrt{n}}} |D_n^w(\boldsymbol{\theta}) - S_n(\boldsymbol{\theta})| \xrightarrow{\mathcal{P}} 0 .$$

The proof is complete. □

Prior to stating the assumptions needed for the asymptotic normality of $\widehat{\widehat{\boldsymbol{\theta}}}_n$, we introduce some helpful notation. Let $y_i^* = y_i - f_i(\boldsymbol{\theta}_0) + \{\nabla f_i(\boldsymbol{\theta}_0)\}^T \boldsymbol{\theta}_0$, $\mathbf{x}_i^* = \nabla f_i(\boldsymbol{\theta}_0)$, and $\mathbf{X}^*$ be the $n \times p$ matrix with $\mathbf{x}_i^*$ as its ith row. Define

$$B_{ij}(t) = E[w_{ij}(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}) I(0 < y_i^* - y_j^* < t) .$$

$$\gamma_{ij} = \frac{\dot{B}_{ij}(0)}{E[w_{ij}(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)})]} .$$

$$C_n(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}) = \sum_{i<j} \gamma_{ij} w_{ij}(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)})(\mathbf{x}_j^* - \mathbf{x}_i^*)(\mathbf{x}_j^* - \mathbf{x}_i^*)^T .$$

$$U_i = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_j^* - \mathbf{x}_i^*) E[w_{ij}(\widehat{\boldsymbol{\theta}}_{V,n}^{(0)}) \{I(y_j^* - y_i^*) - I(y_i^* - y_j^*)\} | y_i^*] .$$

where $\dot{B}$ denotes the derivative of B. Further let A_n be the symmetric $n \times n$ matrix with off-diagonal elements $a_{ij} = -\gamma_{ij} w_{ij}(\widehat{\boldsymbol{\theta}}^{(0)}_{V,n})$ and diagonal elements $a_{ii} = \sum_{k \neq i} \gamma_{ik} w_{ik}(\widehat{\boldsymbol{\theta}}^{(0)}_{V,n})$.

We need the following assumptions.

CN1: There exists a $p \times p$ matrix $C = C(\boldsymbol{\theta}_0)$ such that $n^{-2} C_n(\widehat{\boldsymbol{\theta}}^{(0)}_{V,n}) \xrightarrow{P} C$.

CN2: There exists a $p \times p$ matrix V^* such that $n^{-1} \sum_{i=1}^{n} U_i \rightarrow V^*$.

CN3: $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(0)}_{V,n} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \Xi)$ where Ξ is positive definite.

The following is essentially Theorem 2 of Chang et. al. (1999).

Theorem 5.3.3. *Under WC1-WC5, CN1-CN3, WN2, and N1-N3 of Chapter III,*

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (1/4) C^{-1} V^* C^{-1}) .$$

The result given below is a trivial consequence of Lemma 5.3.2 and Theorem 5.3.3.

Theorem 5.3.4. *Under model* (1.1), *WC1-WC5, CN1-CN3, WN2, and N1-N3 of Chapter III,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{V,n} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (1/4) C^{-1} V^* C^{-1}) .$$

The reader is cautioned that CN1 - CN3 rely heavily on the distributional properties of the initial estimator and it might be difficult to establish CN3 without assuming WN1 or even $E(\varepsilon_1^2) < \infty$. For example, if we intend to use the

LS estimator as an initial estimator, it is more economical to use Theorem 5.3.2 rather than Theorem 5.3.4 since it is assumed that $E(\varepsilon_1^2) < \infty$ in LS asymptotics. Wang (1995) imposes the same assumption in establishing the asymptotic normality of the nonlinear L_1 estimator.

CHAPTER VI

NUMERICAL EXAMPLES AND A SIMULATION STUDY

6.1 Numerical Examples

Although our focus was in developing the asymptotic theory of rank regression for nonlinear models, we consider a few examples that demonstrate the robustness and efficiency properties of the rank estimators in comparison to the least squares(LS) estimator in practical situations. Since most data contain contamination, due to either the faulty nature of the mechanism which produces the data or human error in handling the data, the use of procedures such as the ones developed in this study becomes one of the ways of making sensible inference.

In Chapter I we have seen that the LS estimate is very sensitive to outlying observations. In this chapter we consider more examples depicting this fact and providing one possible remedy. For illustration purposes we will focus on the Wilcoxon estimator given in Chapter III and show that it is a robust alternative to LS. All our estimates are computed using the package RGLM of Kapenga et. al. (1995).

Example 6.1.1 *(Chwirut's data).* These data are taken from the ultrasonic block reference study by Chwirut (1979). The response variable is ultrasonic response and the predictor variable is metal distance. The study involved 214 observations.

The model under consideration is,

$$f_i(\boldsymbol{\theta}) \equiv f(x_i; \theta_1, \theta_2, \theta_3) \equiv \frac{\exp[-\theta_1 x_i]}{\theta_2 + \theta_3 x}. \quad i = 1, \ldots, 214 .$$

Both the Wilcoxon and LS were fitted to the data. Figure 3 displays the results. In the case of the original data, both models performed very similarly. For robustness considerations, we introduced a gross outlier in the response space. The models were fitted once again. From the plot of the fitted models and residual plots, it is clear that the Wilcoxon model performs dramatically better than its LS counterpart. The LS fit follows the "Archemedian lever" principle; that is, put a point far enough out in the response space and the LS fit will go right through it. This is, however, not true in the case of the Wilcoxon estimator which stayed unchanged.

Example 6.1.2 *(Lanczos' data).* In this example we consider a generated data set given in Lanczos (1956). Twenty four observations were generated to 5-digits of accuracy using $f(x) = 0.0951\exp(-x) + 0.8607\exp(-3x) + 1.5576\exp(-5x)$. Naturally the model we consider is,

$$f_i(\boldsymbol{\theta}) = \theta_1 \exp(-\theta_2 x_i) + \theta_3 \exp(-\theta_4 x_i) + \theta_5 \exp(-\theta_6 x_i), \quad i = 1, \ldots, 24.$$

Just as in Example 6.1.1 we fitted both models with and without an outlier present. This time the outlier introduced does not deviate much from the form of the model as can be seen in Figure 4. The effect of the outlier is clearly seen in the LS residual plot which developed a wave-like pattern. LS followed the outlier;

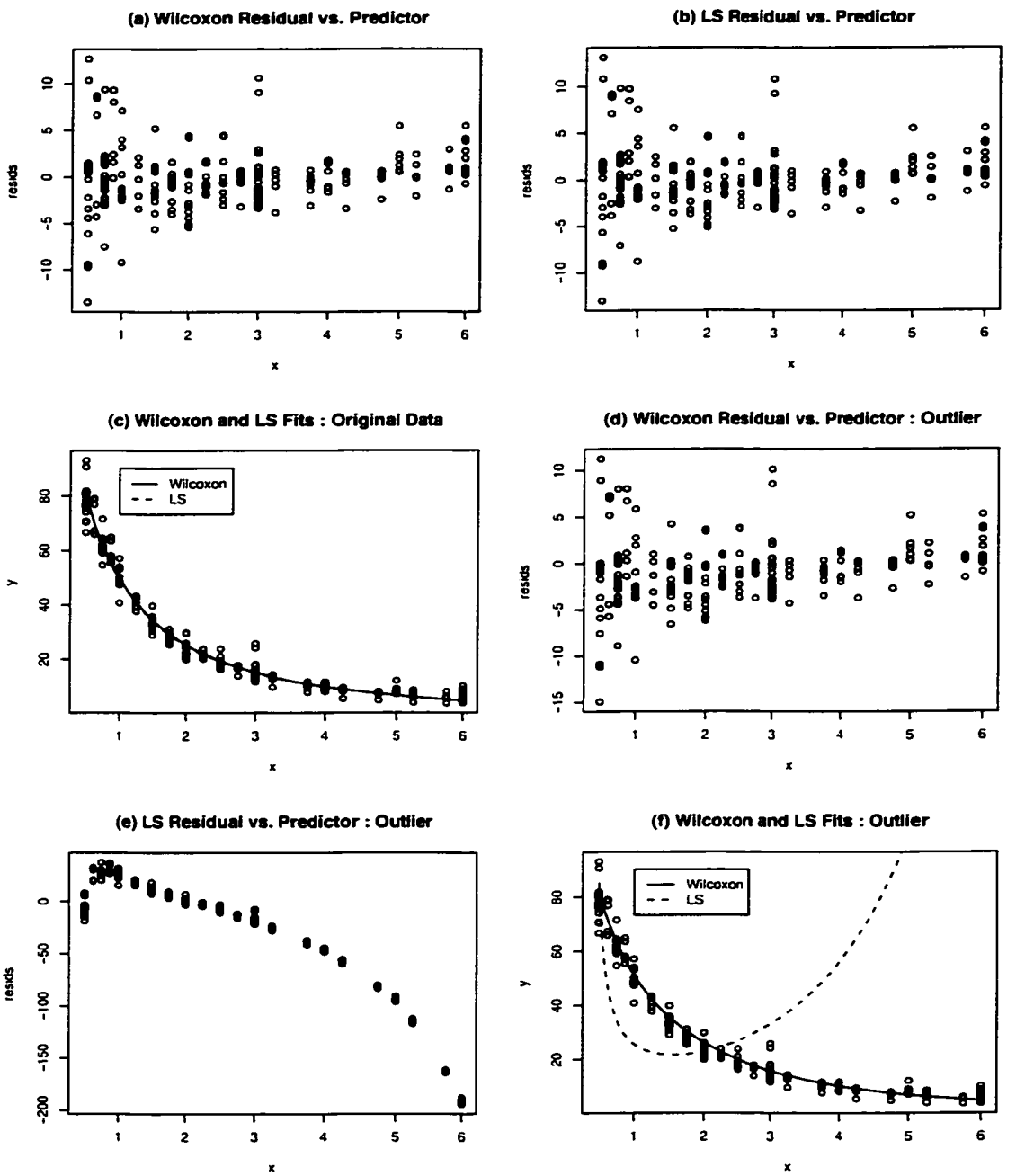

Figure 3. Analysis of Chwirut's data

but since the estimation is done under smoothness and shape restrictions imposed by the model, points in the neighborhood of the outlier will also be affected to some degree. This produces the wave in the residual plot.

6.2 A Simulation Study

It is well known (see for example Hettmansperger and McKean (1998)) that the ARE of the Wilcoxon estimator relative to the LS estimator is about 95.5% when the errors are normally distributed with mean 0 and variance 1. When the distribution of the errors has a heavier tail than the tails of the standard normal distribution the value of the ARE rises substantially. A natural question to ask is "Does this phenomenon hold in nonlinear models?". The answer is affirmative.

We start by defining

$$CN(\gamma, \eta) \equiv (1-\gamma)N(0,1) + \gamma N(0,\eta), \quad .0 \leq \gamma \leq 1 .$$

where $\eta > 0$, to be the contaminated normal distribution. In this case the contaminating distribution is also normal but with a variance different from 1.

Taking the ratio of the asymptotic variances of the LS estimator and the Wilcoxon estimator and applying simple algebra shows that the asymptotic relative efficiency of the Wilcoxon estimator relative to the LS estimator when the errors come from the $CN(\gamma, \eta)$ distribution is given by,

$$\text{ARE}(\gamma, \eta) = 12\Big[\frac{(1-\gamma)^2}{2\sqrt{\pi}} + \frac{\gamma^2}{2\sqrt{\eta\pi}} + \frac{\sqrt{2}\gamma(1-\gamma)}{\sqrt{\pi(\eta+1)}}\Big]^2 [(1-\gamma) + \gamma\eta] .$$

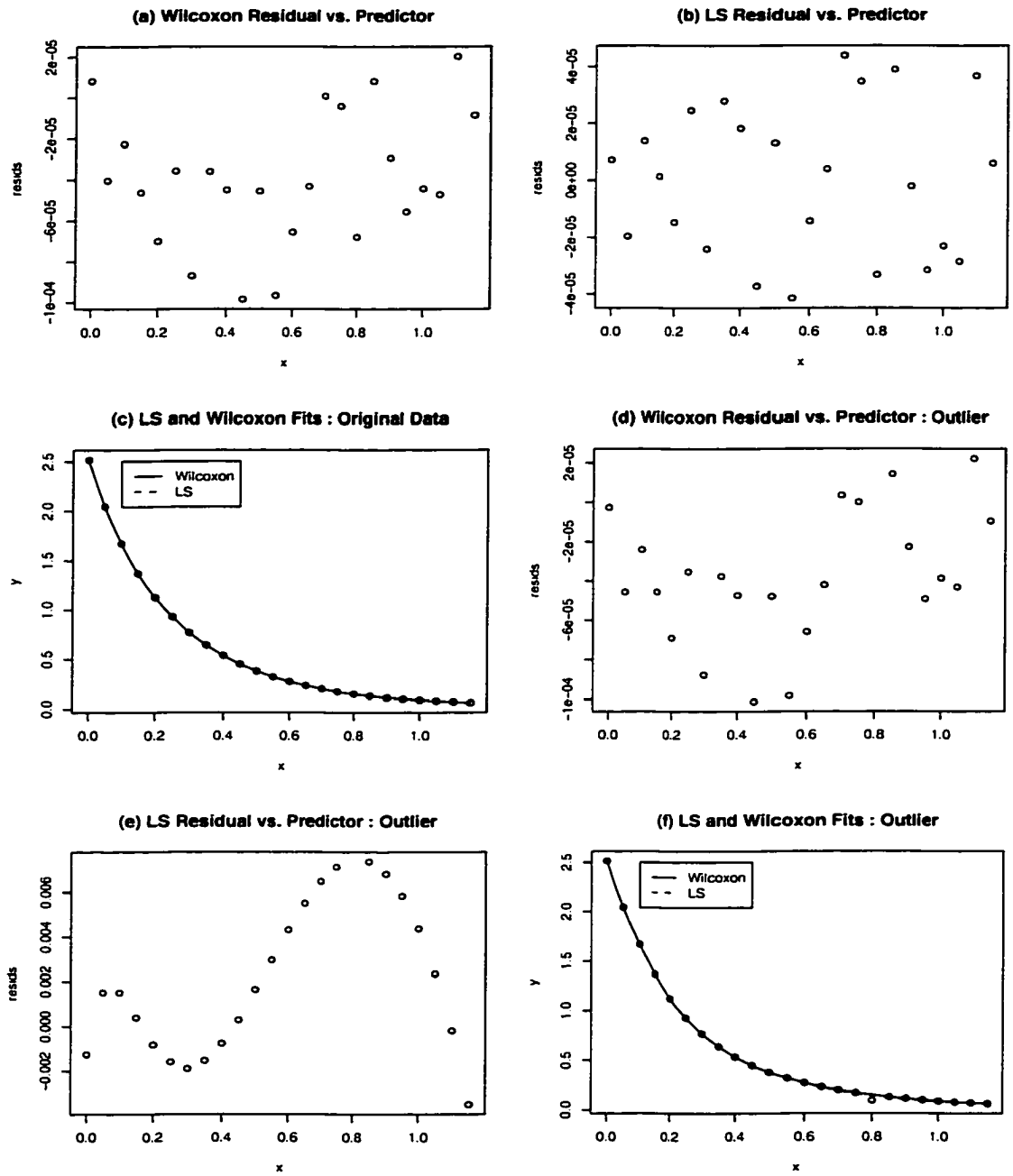

Figure 4. Analysis of Lanczos' data

It is easy to see that $\mathrm{ARE}(0, \eta) = \mathrm{ARE}(1, \eta) = \mathrm{ARE}(\gamma, 1) = 3/\pi$. Furthermore, $\mathrm{ARE}(\gamma, \eta)$ is increasing in both its arguments. So, either an increase in contamination or an increase in the variance increases the ARE. Thus if the error distribution is standard normal then the Wilcoxon estimator is about 95.5% as efficient as the LS estimator.

In order to investigate the efficiency of the Wilcoxon estimator relative to the LS estimator, we consider the function,

$$f_i(\theta) = \exp(x_i\theta), \quad i = 1, \ldots, n\,.$$

This functional form is then used to generate a vector of response by fixing $\theta = \log(2)$ and adding random errors as,

$$y_i = \exp(x_i \log(2)) + \varepsilon_i, \quad x_i = 1, \ldots, n\,,$$

where x_i are uniformly distributed over the interval $(0, 5)$ and ε_i are sampled from, $CN(\gamma, \eta)$.

We performed 1000 repetitions at $n = 20$ and obtained LS and Wilcoxon fits using the algorithm given by Sievers and Abebe (2002). The finite sample relative efficiency (RE) is then taken to be the ratio of the bootstrap variance of the LS fit to that of the Wilcoxon fit. The estimated values of the relative efficiency are given in Table 2. One can observe that the estimated values of RE are in a close proximity of the true ARE values.

Table 2

Estimated relative efficiencies of Wilcoxon relative to LS

	RE($\eta = 3$)	ARE(γ,$\eta = 3$)	RE($\eta = 10$)	ARE(γ,$\eta = 10$)
$\gamma = 0.00$	0.957	0.955	0.960	0.955
$\gamma = 0.01$	1.019	1.009	1.826	1.836
$\gamma = 0.05$	1.193	1.196	4.796	4.769
$\gamma = 0.10$	1.363	1.373	7.399	7.280
$\gamma = 0.15$	1.479	1.497	8.695	8.757
$\gamma = 0.20$	1.558	1.575	9.193	9.430

CHAPTER VII

CONCLUSIONS

7.1 Concluding Remarks

The study considered the general regression model.

$$y_i = f_i(\boldsymbol{\theta}_0) + \varepsilon_i, \quad i = 1, \ldots, n ,$$

where each f_i are known real valued functions defined on a compact space $\boldsymbol{\Theta}$ and ε_i are random errors assumed to be independent and identically distributed. In some cases, instead of several functions f_i, we only have one function f taking several inputs. In such a case the model is represented by

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \varepsilon_i, \quad i = 1, \ldots, n ,$$

where $\mathbf{x}_i \in \mathcal{X} \subset \Re^p$ are the independent variables. Estimation is done by minimizing some distance between y and the expectation surface $\mathcal{S}(\boldsymbol{\Theta})$, the space traced by $f(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ varies over $\boldsymbol{\Theta}$. The most prominent problem encountered when minimizing these distances is that there may be suboptimal minima, the abundance of which depends on the curvature of the space $\mathcal{S}$ and the length of the shortest distance of y from $\mathcal{S}$. Most of the popular distances will be nonconvex due to the nonlinear nature of f.

The main purpose of the study was to establish the asymptotic theory of some estimators of $\boldsymbol{\theta}_0$ obtained by minimizing rank-based dispersion functions. Besides making our proofs economical, the compactness of the parameter space, $\boldsymbol{\Theta}$, is needed to ensure the existence of the estimators. Furthermore some smoothness and identifiability conditions need to be imposed on the space $\mathcal{S}$ to be able to obtain the asymptotic properties of the estimators.

In Chapter II, we developed some tools to aid in developing the asymptotic theory of rank estimators in nonlinear models. We gave a result on convergence of probability measures on compact spaces that was used in proving the consistency of our estimators. We also developed an asymptotic bound on the distance between two minimizers in terms of the distance between the functions minimized in the spirit of Hjort and Pollard (1993) and extending the work of Jaeckel (1972). In our setting, these two results give us the probabilistic tools needed for developing the asymptotic theory of estimators.

In Chapter III, we considered the Wilcoxon estimator which is defined as any value of $\boldsymbol{\theta}$ which minimizes

$$\sum_{i<j} |[y_i - f_i(\boldsymbol{\theta})] - [y_j - f_j(\boldsymbol{\theta})]| .$$

The existence, consistency, and asymptotic normality of the Wilcoxon estimator were obtained under some regularity conditions on the error distribution and f similar to those assumed in L_1 and least squares estimation procedures. The influence function, although bounded in the response space, was shown to be

unbounded if the tangent plane at $\boldsymbol{\theta}_0$ of the expectation surface is unbounded.

In Chapter IV, we studied the generalized signed-rank dispersion function given by

$$\sum_{i=1}^{n} a_n(i)\rho(|z(\boldsymbol{\theta})|_{(i)}) ,$$

where $|z(\boldsymbol{\theta})|_{(i)}$ is the ith ordered value among $|y_1 - f_1(\boldsymbol{\theta})|, \cdots, |y_n - f_n(\boldsymbol{\theta})|$ and ρ is an increasing function defined on $\Re^+$. The scores, a_n, are usually picked using a score generating function, φ^+, which has at most a finite number of discontinuities. One can easily come up with a number of dispersion functions by changing φ^+ and ρ. Examples are least squares and L_1 dispersion functions including their trimmed variations. The conditions for the strong consistency of the minimizer of the generalized signed-rank dispersion were obtained by placing ρ and φ^+ in conjugate spaces. In most cases, these conditions were found to be weaker than those commonly used in practice. One can obtain high breakdown estimates by trimming the score generating function to zero the influence of large residuals. The breakdown point of the generalized signed-rank estimator is shown to be at least equal to the amount trimmed off the top of φ^+.

In Chapter V, we generalized the results of Chapter III by considering a weighted Wilcoxon dispersion given by

$$\sum_{i<j} w_{ij}(\widehat{\boldsymbol{\theta}}_n^{(0)})|[y_i - f_i(\boldsymbol{\theta})] - [y_j - f_j(\boldsymbol{\theta})]| ,$$

where w_{ij} are weight functions and $\widehat{\boldsymbol{\theta}}_n^{(0)}$ is an initial estimator that is weakly consistent. The asymptotic properties of the weighted Wilcoxon estimator were

obtained under conditions equivalent to those used in Chapter III in addition to smoothness and boundedness conditions on w_{ij}. As in linear models (see Chang et. al. (1999)), the weighted Wilcoxon could potentially produce high breakdown nonlinear regression estimators. This conjecture needs further investigation.

The examples given in Chapter VI demonstrate the ability of rank-based estimators to extract valuable information from the data in the presence of outliers when the least squares method fails to do so. This is further testimony that rank-based estimates provide a robust alternative to least squares and need to be included in our data analysis protocols. Besides its robustness properties, our simulation study shows evidence that the Wilcoxon estimator of nonlinear regression coefficients is a highly efficient estimator in comparison to the least squares estimator.

7.2 Future Research Directions

We will start by considering higher generalizations of the nonlinear model considered in (1.1) by allowing stochastic dependence among the f_i.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ be a filtered probability space. Let Θ be a compact subset of $\Re^p$. Consider the general stochastic regression model

$$y_t = f_t(\boldsymbol{\theta}) + \varepsilon_t, \tag{7.1}$$

where y_t are $\mathcal{F}_t$-measurable, $\{\varepsilon_t\}$ is a martingale difference sequence with respect

to the increasing sequence of sub-σ-fields $\{\mathcal{F}_t\}$ such that

$$\sup_t E(\varepsilon_t^2|\mathcal{F}_{t-1}) < \infty \tag{7.2}$$

and $f_t(\boldsymbol{\theta})$ is a $\mathcal{F}_{t-1}$-measurable real valued function of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

In addition to the usual regression models, the model given in (7.1) includes several interesting models such as linear and nonlinear time series models. An example is the nonlinear ARMA(p, q) model given by

$$y_t = h(\varepsilon_{t-q}, \ldots, \varepsilon_{t-1}, y_{t-p}, \ldots y_{t-1}) + \varepsilon_t, \tag{7.3}$$

where h is a real-valued function defined on $\Re^{p+q}$. The function f in (7.1) may also depend on covariates in addition to past y's. An example of such model is the nonlinear autoregressive model with exogenous inputs (NARX) given by (see Lai (1994)),

$$y_t = f(y_{t-1}, \ldots y_{n-p}, x_{n-d}, \ldots, x_{n-d-q}; \boldsymbol{\theta}) + \varepsilon_t\,, \tag{7.4}$$

where $d \geq 1$ is the delay and x_i is the ith stage input.

Denote by $\boldsymbol{\theta}_0$ the unknown true value of $\boldsymbol{\theta}$ satisfying $E(y_t|\mathcal{F}_{t-1}) = f_t(\boldsymbol{\theta}_0)$. Lai and Wei (1982) have given sufficient conditions for the strong consistency of the least squares estimator in the case where f is a stochastic linear model. Lai (1994) gave the consistency of the least squares estimator of nonlinear stochastic regression coefficients under identifiability and differentiability (smoothness) conditions imposed on f. These conditions allowed him to place f in a suitably chosen Hilbert space where martingale results can be used to establish the con-

sistency. Recently, by using a theorem of Andrews (1987). Skouras (2000) was able to show the strong consistency of the LS estimator under an assumption of Lipschitz continuity of f in place of Lai's differentiability assumption.

Define the Wilcoxon estimator as any value of $\boldsymbol{\theta} \in \Theta$ which minimizes

$$D_T(\boldsymbol{\theta}) \equiv \sum_{s<t} w_{ij} \left| [y_t - f_t(\boldsymbol{\theta})] - [y_s - f_s(\boldsymbol{\theta})] \right| . \tag{7.5}$$

Denote this value by $\widehat{\boldsymbol{\theta}}_T$.

Terpstra et. al. (2000) considered autoregressive linear models and used (7.5) to obtain highly efficient estimators of the model coefficients. So our first future research direction is to establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}_T$ under some regularity conditions on w_{ij} and f. Again the approach taken by Andrews (1987) of showing pointwise convergence and imposing a Lipschitz-smoothness condition to obtain uniform laws of large numbers seems to be a promising way to prove the consistency of $\widehat{\boldsymbol{\theta}}_T$.

Finally, consider a linear model where the response variable is observed only when it is above a certain threshold. Without loss of generality assume that this threshold is 0 and write the model as

$$y_i = \max(\mathbf{x}_i^T \boldsymbol{\theta} + \varepsilon_i, 0) , \quad i = 1, \cdots , n .$$

General R estimates for such models were obtained by Lai and Ying (1991). Under the assumption that the true errors are symmetric, Powell (1984) showed that the

L_1 estimator is the minimizer of

$$\sum_{i=1}^{n} |y_i - \max(\mathbf{x}_i^T \boldsymbol{\theta}, 0)| \, .$$

which is the L_1 dispersion function corresponding to the nonlinear regression problem $y_i = \max(\mathbf{x}_i^T \boldsymbol{\theta}, 0) + \varepsilon_i$. Similar symmetry conditions on the errors can be imposed to obtain the dispersion function of the signed-rank estimation procedure. One may wish to investigate the asymptotic properties of the signed-rank estimator.

REFERENCES

Andrews. D. W. K. (1987). "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers". *Econometrica. 55*(6). 1465–1471.

Bhattacharyya. B. B.. Otsuka. Y.. & Richardson. G. D. (1992). "Strong Consistency in Nonlinear Regression with Multiplicative Error". *Communications in Statistics. Theory and Methods. 21*(10). 2825–2831.

Chang. W. H.. McKean. J. W.. Naranjo. J. D.. & Sheather. S. J. (1999). "High-Breakdown Rank Regression". *Journal of the American Statistical Association. 94*(445). 205–219.

Chwirut. D. J. (1979). "Recent Improvements to the ASTM-Type Ultrasonic Reference Block System". Research Report NBSIR 79-1742. National Bureau of Standards. Washington. DC.

Dixon. S. L. & McKean. J. W. (1996). Rank-based analysis of the heteroscedastic linear model. *Journal of the American Statistical Association. 91*(434). 699–712.

Donoho. D. & Huber. P. J. (1983). "The Notion of Breakdown Point". In *A Festschrift for Erich L. Lehmann* (pp. 157–184). Belmont. CA: Wadsworth.

Doob. J. L. (1994). *Measure Theory.* New York: Springer-Verlag.

Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics.* New York: John Wiley & Sons Inc.

Fristedt, B. & Gray, L. (1997). *A Modern Approach to Probability Theory.* Boston, MA: Birkhäuser Boston Inc.

Gonin, R. & Money, A. H. (1985). "Nonlinear L_p-Norm Estimation. I. On the Choice of the Exponent, p, where the Errors are Additive". *Communications in Statistics. A. Theory and Methods, 14*(4), 827–840.

Helmers, R. (1977). "A Strong Law of Large Numbers for Linear Combinations of Order Statistics". Technical Report SW 50/77, Mathematisch Centrum, Amsterdam.

Hettmansperger, T. P. & McKean, J. W. (1998). *Robust Nonparametric Statistical Methods.* London: Edward Arnold.

Hettmansperger, T. P. & Sheather, S. J. (1992). "A Cautionary Note on the Method of Least Median Squares". *The American Statistician, 46*(2), 79–83.

Hjort, N. & Pollard, D. (1993). "Asymptotics for Minimizers of Convex Processes". Technical Report 93may-1, Yale University, Department of Statistics.

Hössjer, O. (1994). "Rank-Based Estimates in the Linear Model with High Break-

down Point". *Journal of the American Statistical Association*, *89*(425), 149–158.

Jaeckel, L. A. (1972). "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals". *The Annals of Mathematical Statistics*, *43*, 1449–1458.

Jennrich, R. I. (1969). "Asymptotic Properties of Non-Linear Least Squares Estimators". *The Annals of Mathematical Statistics*, *40*, 633–643.

Jurečková, J. (1969). "Asymptotic Linearity of a Rank Statistic in Regression Parameter". *The Annals of Mathematical Statistics*, *40*, 1889–1900.

Jurečková, J. (1971). "Nonparametric Estimate of Regression Coefficients". *The Annals of Mathematical Statistics*, *42*, 1328–1338.

Kapenga, J., McKean, J. W., & Vidmar, T. J. (1995). "RGLM: Users Manual". Technical Report 90, Western Michigan University, Department of Mathematics and Statistics.

Lai, T. L. (1994). "Asymptotic Properties of Nonlinear Least Squares Estimates in Stochastic Regression Models". *The Annals of Statistics*, *22*(4), 1917–1930.

Lai, T. L. & Wei, C. Z. (1982). "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems". *The Annals of Statistics*, *10*(1), 154–166.

Lai, T. L. & Ying, Z. (1991). "Rank Regression Methods for Left-Truncated and Right-Censored Data". *The Annals of Statistics*, *19*(2), 531–556.

Lanczos, C. (1956). *Applied Analysis*. Englewood Cliffs, N. J.: Prentice Hall Inc.

Landers, D. (1968). *Existing Und Konsistenz von Maximum Likelihood Schatzern*. PhD thesis, University of Cologne.

Lehmann, E. L. (1997). *Theory of Point Estimation*. New York: Springer-Verlag. Reprint of the 1983 original.

Malinvaud, E. (1970). "The Consistency of Nonlinear Regressions". *The Annals of Mathematical Statistics*, *41*, 956–969.

McKean, J. W. & Hettmansperger, T. P. (1978). "A Robust Analysis of the General Linear Model Based on one Step *R*-Estimates". *Biometrika*, *65*(3), 571–579.

McKean, J. W. & Schrader, R. M. (1980). "The Geometry of Robust Procedures in Linear Models". *Journal of the Royal Statistical Society. Series B. Methodological*, *42*(3), 366–371.

Naranjo, J. D. & Hettmansperger, T. P. (1994). "Bounded Influence Rank Regression". *Journal of the Royal Statistical Society. Series B. Methodological*, *56*(1), 209–220.

Oberhofer, W. (1982). "The Consistency of Nonlinear Regression Minimizing the L_1-Norm". *The Annals of Statistics*, *10*(1), 316–319.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory.* New York: The Clarendon Press Oxford University Press. Sequences of independent random variables. Oxford Science Publications.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability.* Cambridge: Cambridge University Press.

Powell, J. L. (1984). "Least Absolute Deviations Estimation for the Censored Regression Model". *Journal of Econometrics*. *25*(3), 303–325.

Prakasa Rao, B. L. S. (1987). *Asymptotic Theory of Statistical Inference.* New York: John Wiley & Sons Inc.

Pronzato, L. & Walter, E. (2001). "Estimating Suboptimal Local Minimizers in Nonlinear Parameter Estimation". *Technometrics*. *43*(4), 434–442.

Richardson, G. D. & Bhattacharyya, B. B. (1986). "Consistent Estimators in Nonlinear Regression for a Noncompact Parameter Space". *The Annals of Statistics*, *14*(4), 1591–1596.

Rousseeuw, P. J. (1983). "Regression Techniques with High Breakdown Point". *The Institute of Mathematical Statistics Bulletin*, (12), 155.

Rousseeuw, P. J. (1984). "Least Median of Squares Regression". *Journal of the American Statistical Association, 79*(388), 871–880.

Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear Regression.* New York: John Wiley & Sons Inc.

Sen, P. K. (1978). "An Invariance Principle for Linear Combinations of Order Statistics". *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 42*(4), 327–340.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

Sievers, G. L. (1983). "A Weighted Dispersion Function for Estimation in Linear Models". *Communications in Statistics. A. Theory and Methods. 12*(10), 1161–1179.

Sievers, G. L. & Abebe, A. (2002). "Rank Estimation of Regression Coefficients Using Iterated Reweighted Least Squares". *Submitted.*

Skouras, K. (2000). "Strong Consistency in Nonlinear Stochastic Regression Models". *The Annals of Statistics, 28*(3), 871–879.

St. Laurent, R. T. & Cook, R. D. (1993). "Leverage, Local Influence and Curvature in Nonlinear Regression". *Biometrika, 80*(1), 99–106.

Strasser, H. (1973). "On Bayes Estimates". *Journal of Multivariate Analysis, 3.* 293–310.

Stromberg, A. J. (1995). "Consistency of the Least Median of Squares Estimator in Nonlinear Regression". *Communications in Statistics. Theory and Methods. 24*(8). 1971–1984.

Stromberg, A. J. & Ruppert, D. (1992). "Breakdown in Nonlinear Regression". *Journal of the American Statistical Association, 87*(420), 991–997.

Terpstra, J. T., McKean, J. W., & Naranjo, J. D. (2000). Highly Efficient Weighted Wilcoxon Estimates for Autoregression". *Statistics, 35*(1), 45–80.

van Zwet, W. R. (1980). "A Strong Law for Linear Functions of Order Statistics". *The Annals of Probability, 8*(5). 986–990.

Wang, J. D. (1995). "Asymptotic Normality of L_1-Estimators in Nonlinear Regression". *Journal of Multivariate Analysis, 54*(2), 227–238.

Wellner, J. A. (1977). "A Glivenko-Cantelli Theorem and Strong Laws of Large Numbers for Functions of Order Statistics". *The Annals of Statististics. 5*(3), 473–480.

Willard, S. (1970). *General Topology.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Williams, D. (1991). *Probability with Martingales.* Cambridge: Cambridge University Press.

Wu, C.-F. (1981). "Asymptotic Theory of Nonlinear Least Squares Estimation". *The Annals of Statistics, 9*(3), 501–513.