Rank-Based Regression for Nonlinear and Missing Response Models

by

Huybrechts Frazier Achard Bindele

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 4, 2012

Keywords: Asymptotic normality, bounded influence, signed-rank norm, Sobolev space, strong consistency, missing at random.

Approved by

Asheber Abebe, Chair, Associate Professor of Mathematics & Statistics
Geraldo S. de Souza, Professor of Mathematics & Statistics
Peng Zeng, Associate Professor of Mathematics & Statistics
Mark D. Carpenter, Professor of Mathematics & Statistics

Abstract

This dissertation is mainly concerned with the rank-based estimation of model parameters in complex regression models: a general nonlinear regression model and a semi-parametric regression model with missing responses. For the estimation of nonlinear regression parameters, we consider weighted generalized-signed-rank estimators. The generalization allows us to study rank estimators as well as popular estimators such as the least squares and least absolute deviations estimators. However, the generalization by itself does not give bounded influence estimators. Added weights provide estimators with bounded influence function. We establish conditions needed for the consistency and asymptotic normality of the proposed estimator and discuss how weight functions can be chosen to achieve bounded influence function of the estimator. Real life examples and Monte Carlo simulation experiments demonstrate that the proposed estimator is robust, efficient, and useful in detecting outliers in nonlinear regression. For the estimation of the linear regression parameter of a semi-parametric model with missing response, we propose imputed rank estimators under simple imputation and imputation by inverse probability. It is shown that these rank estimators have favorable asymptotic properties. Moreover, it is demonstrated that the rank estimators perform better than the classical least squares estimator under heavy tailed error distributions and cases containing contamination while they are generally comparable to the least squares estimator under normal error. Moreover, rank estimators with inverse probability imputation are superior than their least squares counterpart when the proportion of missing data is large. This makes rank estimation extremely appealing for situations where we encounter high rates of missing information.

Acknowledgments

Let me start by thanking God for the strength, the health and the knowledge given to me throughout all these years and simply throughout my life. Without all your blessings, I could never have accomplished this long journey.

This research project would not have been possible without the support of many people. I wish to express my sincere gratitude to my supervisor, Dr. Ash Abebe who was abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to the members of the supervisory committee, Drs. Geraldo de Souza, Peng Zeng, and Mark Carpenter without whose knowledge and assistance this study would not have been successful.

Big dedication of this dissertation to my parents Philippe Bindele and Suzanne Mbeke who have always been supportive and have made a lot of sacrifices for the great success of my life.

Special thanks to my grand mother Therese Mayena and his husband Andre Moumpangou, who, even though did not live long enough to see this achievement, represent the key point of my success by providing all necessary advices and encouragement to my education life. Another special thanks to my uncle Cyrille Ngayi Mvouembe who always stands up for me for any educational purposes.

I would like also to express my profound gratitude to the department Chair and the graduate Chair of the Mathematics and Statistics department Dr. Michel Smith, Chris Rodger and Paul Schmidt for all their supports and encouragements during my graduate student life at Auburn University. At the same time, I would like to thank our department staff: Lori Bell, Gwen Kirk and Carolyn Donegan for their tremendous help provided when needed. I wishes to express my gratitude to the following:

Prof. Charles Chidume whose supports, advices and encouragements has made my journey to the US possible.

All Professors who taught me during the course work for the marvelous job done, particularly, Drs. Wanxian Shen, Jersy Szulga, Erkan Nane, Geraldo de Souza, Asheber Abebe, Mark Carpenter, Trevor Park and Peng Zeng. Also all beloved faculty members such as Drs. Greg Harris, Nedret Billor, Overtoun Jenda, Peter Johson, Olav Kallenberg, Amnon Meir, Andreas Bezdek, Tin-Yao Tam, Bertram Zinner, Frank Uhlig and Narendra Govil.

My course mates and friends, Gerald Chidume, Julian Allagan, Fidele Ngwane, Nar Rawal, Bertrand Sedar Ngoma, Guy-Vanie Miankokana, Brice Merlin Nguelifack, Sean O'Neil, Frank Sturn, Serge Phanzu, Jebessa Mijena, Melody Donhere, Dawit Tadesse, Omer Tadjion, Al-Hassem Nayam, Daniel Bongue, Guy Richard Bossoto, Justin Mabiala, Simplice Mananga, Brice Malonda, Justin Moutsouka, Seth Kwame Kermaussuor, Kuena, Gouala Medea, Telemina Gaston, Charles Lebon Mberi Kimpolo, Eddy Kwessi, Rolland Kendou, Merlin Ngachi, Gladisse Ngodo, Moses Tam, Murielle Mahoungou, Aude Didas Massika, Roland Loufouma, Darius Kissala, Fortuné Massamba, Amin Bahmanian, Patrice Bassissa for the good time and fun we have had.

My brothers, sisters and family members, Durand Olivier Ngoyi, Elvis Darel Ngouala Nzaou, Tancraide Malanda Ngouala, Mabelle Ngouala Mboussi, Aude Durgie Nsimba, Nadine Dihoulou, Emilie Kiloula, Michel Ngoyi, Joseph Nzoussi, Joseph Kimbassa, Antoine Yila, Guyen Kitanda Ndolo, Martin Ndolo, Serge Mawa, Daniel Ndamba, Marie Laure Gayi for their support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Backround

The historical approach to fitting linear and nonlinear models of the form:

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

for some generic function f (linear or nonlinear), proceeds by finding coefficient estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ that minimizes the sum of squared errors: $\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}))^2$. Such estimators, known as least squares estimators, are computationally simple and possess general optimality properties. However, the optimality can be lost due to the existence of even a single extreme outlier data point. This problem is seen with the sample mean, $\bar{y}$, which is itself the least squares solution to the model $y_i = \boldsymbol{\beta}_0 + \varepsilon_i$. To overcome this problem, we briefly survey a few approaches that have been taken to develop estimators of the $\boldsymbol{\beta}$ coefficients that are not as easily affected as the least-squares estimators. Two methods have been investigated: one is a generalization of least-squares estimation called M-estimation, and the other is the so-called rank based estimation methods referred to as rank-based regression methods. For each of these two approaches a first method was developed to downweight outlier data points, but was later shown to be susceptible to high leverage points (outliers in the $\mathbf{x}$ space) in regression problems, and newer methods have emerged to address both outlier and leverage problems.

One approach that has been used to lessen the impact of outliers in linear and nonlinear models is to use the least abolute deviation also known as the L_1 regression, that is, finding coefficient estimates $\widehat{\boldsymbol{\beta}}$ that minimize $\sum_{i=1}^{n} |y_i - f(\mathbf{x}_i, \boldsymbol{\beta})|$. A further generalization to this,

was made by Huber in the 1960s. He obatained the so-called M-estimators by minimizing $\sum_{i=1}^{n} \rho\Big(\frac{y_i - f(\mathbf{x}_i,\boldsymbol{\beta})}{\widehat{\sigma}_i}\Big)$, where $\rho(\cdot)$ is a symmetric function and $\widehat{\sigma}_i$ is an estimate of the standard devaition of the errors ε_i. It was shown that these M-estimators have the advantage of downweighting outliers while retaining efficiency when compared to least squares estimators. However, the original M-estimators can be affected by leverage points (outliers in the $\mathbf{x}$ space) in regression problems. A type of M-estimator developed to protect against aoutliers and leverage points, is the least trimmed squares estimator that minimizes $\sum_{i=1}^{k} \big(y_i - f(\mathbf{x}_i,\boldsymbol{\beta})\big)^2_{(i)}$, where $k \leq n$. This estimator ignores the largest $n-k$ residuals. However, the fact that it does not use the entire data results in a loss of efficiency. More recently, Yohai and others have developed extensions of these methods, called MM estimators, that protect against both outliers and leverage points while retaining efficiency.

At the time when Huber and others were developing the theory of M-estimators, methods based on ranking were known as R-estimation and were used for simple problems such as estimating location and scale or making location comparisons for two-sample problems. They were not considered to be generalizable to linear and nonlinear models. Later Jaeckel, Hettmansperger, McKean and others showed that rank-based estimators could also be cast as estimators obtained by minimizing $\sum_{i=1}^{n} a\big(R(z_i(\boldsymbol{\beta}))\big) z_i(\boldsymbol{\beta})$, where $R(z_i(\boldsymbol{\beta}))$ is the rank of $z_i(\boldsymbol{\beta}) = y_i - f(\mathbf{x}_i,\boldsymbol{\beta})$ and $a(\cdot)$ some score function. The rank-based estimators, sometimes called Wilcoxon estimators, can be used for any general linear model, and have been shown to have high efficiency compared to least squares estimators. However, these rank-based estimators, can also be affected by leverage points in regression problems. A weighted Wilcoxon (WW) method were later developed to take care of leverage points and shown to possess highly efficient.

1.2 Contribution

The first part of this Ph.D dissertation is concerned with the study of conditions sufficient for strong consistency of a class of estimators of parameters of nonlinear regression models. The study considers continuous functions depending on a vector of parameters and a set of random regressors. The estimators chosen are minimizers of a generalized form of the signed-rank norm, that is, minimizing $\frac{1}{n}\sum_{i=1}^{n} a_n(i)\rho(|z(\boldsymbol{\beta})|_{(i)})$, where $z_i(\boldsymbol{\beta}) = y_i - f(\mathbf{x}_i, \boldsymbol{\beta})$ and $|z(\boldsymbol{\beta})|_{(i)}$ is the ith ordered value among $|z_1(\boldsymbol{\beta})|, \ldots, |z_n(\boldsymbol{\beta})|$. The function $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is continuous and strictly increasing. The numbers $a_n(i)$ are scores generated as $a_n(i) = \varphi^+(i/(n+1))$, for some bounded score function $\varphi^+ : (0,1) \to \mathbb{R}^+$ that has at most a finite number of discontinuities. The generalization allows to make consistency statements about minimizers of a wide variety of norms including the L_1 and L_2 norms. By implementing trimming, it is shown that high breakdown estimates can be obtained based on the proposed dispersion function.

The second part of this dissertation is motivated by the fact that almost all the methods discussed above fail to take care of outliers and high leverage points in regression problems. The generalization of the signed-rank norm (above) allows us to include popular estimators such as the least squares and least absolute deviations estimators but by itself does not give bounded influence estimators. To address this problem, we considered weighted forms of the generalized signed-rank estimators and measured the robustness of these estimators using their influence functions. Carefully chosen weights result in estimators with bounded influence function. Also conditions needed for the consistency and asymptotic normality of the proposed estimator are established and discussions about how weight functions can be chosen to achieve bounded influence function of the estimator are provided. Real life examples and Monte Carlo simulation experiments demonstrate the robustness and efficiency of the proposed estimator. Finally, an example showing that the weighted signed-rank estimators can be useful to detect outliers in nonlinear regression is provided.

The last part of this thesis, considered a robust estimation for the missing data problem in regression analysis as an application of the weighted signed-rank estimation method introduced above. In this particular case, important work has been done in the estimation of the regression parameters involving the least squares method and eventually the M-estimation method in the literature. As pointed out by many authors, for non-normal error distributions or in the case of uncontrolled designs, the least squares method may not provide suitable estimators. To this end, we consider the linear semi-parametric regression model with missing response at random, that is, $y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(T_i) + \varepsilon_i$, $i = 1 \cdots n$, where the y_i's are missing at random, $\mathbf{x}_i$'s and T_i's are fully observed. Based on the rank objective function, we provide estimators of $\boldsymbol{\beta}$ and g simultaneously, using the usual nonparametric kernel estimation. Also, under some suitable conditions, we show that the obtained estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$-consistent and also satisfies the normal approximation property. To illustrate this case a simulation study is conducted and shows that the rank estimator behaves better than the least squares estimator when dealing with missing response problem.

Chapter 2

On the Consistency of a Class of Nonlinear Regression Estimators

2.1 Introduction

Over the last twenty five years considerable work has been done on robust procedures for linear models. Several classes of robust estimates have been proposed for these models. One such class is the generalized signed-rank class of estimates. This class uses an objective function which depends on the choice of a score function, φ^+. If φ^+ is monotone then the objective 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; see the monograph by Hettmansperger and McKean (1998) for a discussion of this analysis. 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.

Many interesting problems, though, are nonlinear in nature. Traditional procedures based on LS estimation have been used for years. Since these LS procedures for nonlinear models use the Euclidean norm they are as easily interpreted as their linear model counterparts. The asymptotic theory for nonlinear LS has been developed by Jennrich (1969) and Wu (1981), among others. In this chapter, we propose a nonlinear analysis based on the signed-rank objective function. The objective function is a norm if φ^+ is monotone; hence,

the estimates are easily interpretable. We keep our development quite general, though, to include nonlinear estimates based on Hössjer-type estimates also. Hence our estimates include the nonlinear extensions of the signed-rank Wilcoxon estimate and the L_1 estimate as well as the extensions of high breakdown linear model estimates. Thus we offer a rich family of estimates from which to select for nonlinear models.

In Section 2.2 we present our family of estimates for nonlinear models. In Section 2.3, we show that these estimates are strongly consistent under certain assumptions. We discuss these assumptions, contrasting them with assumptions for current existing estimates. The same section contains a general discussion of interesting special cases such as the L_1 and the Wilcoxon. Section 2.4 discusses the conditions needed to achieve positive breakdown of our estimator.

2.2 Definition and Existence

Consider the following general regression model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + e_i, \qquad 1 \leq i \leq n \tag{2.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is a vector of parameters, $\mathbf{x}_i \in \mathbb{X}$ is a vector of independent variables, and f is a real-valued function defined on $\mathbb{X} \times \boldsymbol{\Theta}$. Let $\mathbf{V} = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ be the set of sample data points. Note that $\mathbf{V} \subset \mathbb{V} \equiv \mathbb{R} \times \mathbb{X}$.

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 a continuous function of $\boldsymbol{\theta}$ for each $\mathbf{x} \in \mathbb{X}$ and a measurable function of $\mathbf{x}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

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

$$D_n(\mathbf{V}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} a_n(i) \rho(|z(\boldsymbol{\theta})|_{(i)}) \tag{2.2}$$

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 $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is continuous and strictly increasing. The numbers $a_n(i)$ are scores generated as $a_n(i) = \varphi^+(i/(n+1))$, for some bounded score function $\varphi^+ : (0,1) \to \mathbb{R}^+$ that has at most a finite number of discontinuities.

This estimator will be denoted by $\widehat{\boldsymbol{\theta}}_n$.

Because $D_n(\mathbf{V}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, Lemma 2 of Jennrich (1969) implies the existence of a minimizer of $D_n(\mathbf{V}, \boldsymbol{\theta})$.

We adopt Doob's (1994) convention and denote by L^p, $1 \leq p \leq \infty$, the space of measurable functions $h : (0,1) \to \mathbb{R}$ 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 $\|h\|_\infty \equiv \text{ess sup } |h|$ for $p = \infty$. 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.

2.3 Consistency

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $i = 1, \ldots, n$, assume that $\mathbf{x}_i$ and $e_i = y_i - f(\mathbf{x}_i; \boldsymbol{\theta}_0)$ 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 $\mathbf{x}_1$, e_1 and $|z_1(\boldsymbol{\theta})|$ respectively. Let $\tilde{G}_\theta$ denotes the distribution of $|z(\theta)|$ and we will assume

A1: $P(f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta}_0)) < 1$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$;

A2: for $1 \leq q \leq \infty$, assume there exists a function h such that $|\rho(G_\theta^{-1}(y))| \leq h(y)$, $\forall\ \theta \in \Theta$ with $E[h^q(Y)] < \infty$ and,

A3: G has a density g that is symmetric about 0 and strictly decreasing on $\mathbb{R}^+$.

As usual, we let $a.s.$ convergence, denote almost sure convergence, i.e., pointwise convergence everywhere except for possibly an event in $\mathcal{F}$ of probability 0.

Theorem 1. *Under A1 - A3,* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

Before giving the proof of this theorem, let us discuss some related lemmas and corollaries that will be used in the proof.

Let $\xi_{(1)}, \ldots, \xi_{(n)}$ be order statistics from a sample of n i.i.d. uniform$(0,1)$ random variables. Let $J_n : (0,1) \to \mathbb{R}$, $n = 1, 2, \ldots$ be Lebesgue measurable functions and let $g : (0,1) \to \mathbb{R}$ be a Borel measurable function. Define $g_n(t) \equiv g(\xi_{([nt]+1)})$. In the defining expression for the function $D_n(\mathbf{V}, \boldsymbol{\theta})$, (2.2), let $\tilde{G}_{\boldsymbol{\theta}}$ dnote the cdf of $|z(\boldsymbol{\theta})|$. Then we can express $D_n(\mathbf{V}, \boldsymbol{\theta})$ as

$$D_n(\mathbf{V}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} a_n(i)(\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1})(\xi_{(i)}). \tag{2.3}$$

The following is Corollary 2.1 of van Zwet (1980) in the notation of this paper and is given for completeness.

Lemma 1 (van Zwet)**.** *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 $\int J_n g_n \xrightarrow{a.s} \int Jg$.

For our purposes let $J_n(t) = \sum_{i=1}^{n} \varphi^+(i/(n+1)) I_{((i-1)/n, i/n]}(t)$ for $i = 1, \ldots, n$ where I_A is the indicator of the set A and take $J = \varphi^+$. Notice that J_n is a step function and thus the uniform integrability condition in assumption (ii) of Lemma 1 becomes

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

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

$$\sup_n \|J_n\|_p \equiv \sup_n \left\{ \frac{1}{n} \sum_{i=1}^{n} |\varphi^+(i/(n+1))|^p \right\}^{1/p} < \infty \tag{2.4}$$

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

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

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 (2.4) and $g \in L^q$ all the conditions of Lemma 1 hold. The following corollary is a special case of this result.

Corollary 1. *Let $W_1, \ldots, W_n$ be a random sample from a distribution F with support on $\mathbb{R}^+$. Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ 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*

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

A formal proof of Corollary 1 may be constructed along the lines described in the paragraph preceding it with the function g defined as $\rho \circ F^{-1}$. It will not be included here for the sake of brevity.

Lemma 2. *Under assumptions A1 - A3*

$$D_n(\mathbf{V}, \boldsymbol{\theta}) \xrightarrow{a.s.} \mu(\boldsymbol{\theta}) \qquad \textit{a.e. } \mathbb{V}\textit{, uniformly for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{2.5}$$

where $\mu : \boldsymbol{\Theta} \to \mathbb{R}$ *is a function satisfying*

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0), \tag{2.6}$$

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

Proof. The *a.s.* pointwise convergence of $D_n(\mathbf{V}, \boldsymbol{\theta})$ follows from expression (2.3) and Corollary 1, which also furnishes the function

$$\mu(\boldsymbol{\theta}) \equiv \int (\varphi^+) \ (\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1}) < \infty \ . \tag{2.7}$$

Then under $A1$ - $A3$, Theorem 2 of Jennrich (1969) gives (2.5).

To establish (2.6) we follow a similar strategy as in Hössjer (1994). Under $A1$ and $A3$ for any $s > 0$, for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

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

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 $\mu^* \in \mathbb{R}$ such that $\mu(\boldsymbol{\theta}) > \mu^* > \mu(\boldsymbol{\theta}_0)$. Then it follows from the compactness of $\boldsymbol{\Theta}^*$ that $\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}^*} \mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0)$. □

Lemma 3. *Let* $\{h_n\}$ *be a sequence of continuous functions defined on a compact set* $\boldsymbol{\Theta} \subset \mathbb{R}^p$ *and that converges uniformly to* h. *Then* $\{h_n\}$ *is equicontinuous on* $\boldsymbol{\Theta}$.

Proof. Since $\{h_n\}$ converges uniformly to h, for any $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $|h_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| < \epsilon/3$ for all $n \geq N$. The function h being continuous on a compact set, it

is uniformly continuous. Thus there exists some $\delta > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta$, we have $|h(\boldsymbol{\theta}) - h(\boldsymbol{\theta}')| < \epsilon/3$. Then for all $n \geq N$ and for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta$, we have

$$|h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}')| \leq |h_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| + |h_n(\boldsymbol{\theta}') - h(\boldsymbol{\theta}')| + |h(\boldsymbol{\theta}) - h(\boldsymbol{\theta}')| < \epsilon \, .$$

Also, by uniform continuity of $\{h_n\}$, for any fixed $n \in \{1, \ldots, N-1\}$, there exists a $\delta_n > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta_n$, we have $|h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}')| < \epsilon$. Now set $\delta' = \min\{\delta_1, \ldots, \delta_{N-1}\}$. Then for all $n \in \{1, \ldots, N-1\}$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta'$, we have $|h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}')| < \epsilon$.

Therefore, setting $\Delta = \min\{\delta, \delta'\}$, for all $n \in \mathbb{N}$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \Delta$, we have $|h_n(\boldsymbol{\theta}) - h_n(\boldsymbol{\theta}')| < \epsilon$. □

Proof of Theorem 1. By Lemma 1 of Wu (1981), to establish the consistency of $\widehat{\boldsymbol{\theta}}_n$, it is sufficient to show that

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \big(D_n(\mathbf{V}, \boldsymbol{\theta}) - D_n(\mathbf{V}, \boldsymbol{\theta}_0)\big) > 0 \quad \text{a.s.} \tag{2.8}$$

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

$$\begin{aligned}\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \big(D_n(\mathbf{V}, \boldsymbol{\theta}) - D_n(\mathbf{V}, \boldsymbol{\theta}_0)\big) \geq \liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} A_n(\mathbf{V}, \boldsymbol{\theta}) + \\ \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \liminf_{n\to\infty} C_n(\mathbf{V}, \boldsymbol{\theta}_0),\end{aligned} \tag{2.9}$$

where $A_n(\mathbf{V}, \boldsymbol{\theta}) = D_n(\mathbf{V}, \boldsymbol{\theta}) - \mu(\boldsymbol{\theta})$, $B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta}_0)$, and $C_n(\mathbf{V}, \boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}_0) - D_n(\mathbf{V}, \boldsymbol{\theta}_0)$.

As a result of Corollary 1, $\liminf_{n\to\infty} C_n(\mathbf{V}, \boldsymbol{\theta}_0) = 0$ a.s. Due to Lemma 2 we have $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > 0$. For the statement given in (2.8) to hold, it suffices to show is that $\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} A_n(\mathbf{V}, \boldsymbol{\theta}) = 0$ a.s. Again by Lemma 2, $A_n(\mathbf{V}, \boldsymbol{\theta}) \xrightarrow{a.s.} 0$ uniformly for all

$\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$. Also $A_n(\mathbf{V}, \boldsymbol{\theta})$, being continuous on a compact set $\boldsymbol{\Theta}^*$, is uniformly continuous on $\boldsymbol{\Theta}^*$. Then $A_n(\mathbf{V}, \boldsymbol{\theta})$ is equicontinuous on $\boldsymbol{\Theta}^*$ a.e. $\mathbb{V}$ by Lemma 3. Thus $\forall\ \epsilon > 0$ there exists a $\delta > 0$ such that $\forall\ \boldsymbol{\theta},\ \boldsymbol{\theta}' \in \boldsymbol{\Theta}^*$,

$$\text{if } \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta \text{ then } |A_n(\mathbf{V}, \boldsymbol{\theta}) - A_n(\mathbf{V}, \boldsymbol{\theta}')| < \epsilon, \quad a.e.\ \mathbb{V}, \quad \forall\ n \in \mathbb{N}. \tag{2.10}$$

Let $D_{\boldsymbol{\theta}'} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta\}$, for $\boldsymbol{\theta}' \in \boldsymbol{\Theta}^*$. Then $D_{\boldsymbol{\theta}'}$, $\boldsymbol{\theta}' \in \boldsymbol{\Theta}^*$, forms an open covering of $\boldsymbol{\Theta}^*$. But $\boldsymbol{\Theta}^*$ is compact, hence there is a finite subcovering $D_{\boldsymbol{\theta}'_j}$, $j = 1, \ldots, m$ which covers $\boldsymbol{\Theta}^*$. Let $\boldsymbol{\theta}^*$ be an arbitrary point in $\boldsymbol{\Theta}^*$. Then for some $j = 1, \ldots, m$, $\boldsymbol{\theta}^* \in D_{\boldsymbol{\theta}'_j}$. Hence, $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}'_j\| < \delta$. Thus by (2.10)

$$|A_n(\mathbf{V}, \boldsymbol{\theta}^*) - A_n(\mathbf{V}, \boldsymbol{\theta}'_j)| < \epsilon, \quad a.e.\ \mathbb{V}, \quad \forall\ n \in \mathbb{N}\,.$$

That is,

$$A_n(\mathbf{V}, \boldsymbol{\theta}'_j) - \epsilon < A_n(\mathbf{V}, \boldsymbol{\theta}^*) < A_n(\mathbf{V}, \boldsymbol{\theta}'_j) + \epsilon, \quad a.e.\ \mathbb{V}, \quad \forall\ n \in \mathbb{N}$$

which implies that

$$\min_{1 \le j \le m} A_n(\mathbf{V}, \boldsymbol{\theta}'_j) - \epsilon < A_n(\mathbf{V}, \boldsymbol{\theta}^*) < \max_{1 \le j \le m} A_n(\mathbf{V}, \boldsymbol{\theta}'_j) + \epsilon, \quad a.e.\ \mathbb{V}, \quad \forall\ n \in \mathbb{N}\,.$$

Since, $\boldsymbol{\theta}^*$ is arbitrary, we have

$$\min_{1 \le j \le m} A_n(\mathbf{V}, \boldsymbol{\theta}'_j) - \epsilon < \inf_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} \{A_n(\mathbf{V}, \boldsymbol{\theta}^*)\} < \max_{1 \le j \le m} A_n(\mathbf{V}, \boldsymbol{\theta}'_j) + \epsilon, \quad a.e.\ \mathbb{V}, \quad \forall\ n \in \mathbb{N}\,.$$

Now take lim inf of all three parts as $n \to \infty$. Since the functions min and max are continuous, we have

$$0 - \epsilon \le \liminf_{n \to \infty} \inf_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} \{A_n(\mathbf{V}, \boldsymbol{\theta}^*)\} \le 0 + \epsilon \quad a.s.$$

Since ϵ was arbitrary, we have $\liminf_{n \to \infty} \inf_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*} \{A_n(\mathbf{V}, \boldsymbol{\theta}^*)\} = 0 \quad a.s.$ The proof is complete. □

Remark 1. *Assumption A1 is a very weak condition needed for* $\boldsymbol{\theta}_0$ *to be identified. The linear version of A1 was given by Hössjer (1994) as* $P(|\boldsymbol{\theta}'\mathbf{x}| = 0) < 1$ *under the assumption that* $\boldsymbol{\theta}_0 = 0$.

Remark 2. *Since* $\|\varphi^+\|_p < \infty$ *for* p *such that* $1/p + 1/q = 1$, *then A2 puts* h *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*

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

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

Remark 3. *Condition A3 admits a wide variety of error distributions examples of which are the normal, double exponential and Cauchy distributions with location parameter equal to 0.*

Some Corollaries

Next some special cases of interest are considered. We consider the L_1, least squares, signed-rank Wilcoxon, and their trimmed variations. All these cases involve a convex ρ and hence Remark 2 is directly applicable. Trimming is implemented by "chopping-off" the ends of the score generating function, φ^+ [cf Hössjer (1994)]. The proofs follow from Theorem 1 in a straightforward manner.

Least Squares, Least Trimmed Squares

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 (2.2) is the least squares dispersion function. If $0 < \alpha < \beta < 1$, 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 by taking $p = q = 2$ in Theorem 1.

Corollary 2. *If*

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

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

B3: G *has a density* g *that is symmetric about 0 and strictly decreasing on* $\mathbb{R}^+$,

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

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

J3: $E(e) = 0$.

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. Proceeding to compare B1 and J1, 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. The converse is also immediate. Hence our assumptions reduce to the assumptions of Jennrich (1969) in the case of least squares.

For linear models, the consistency of the least trimmed squares estimator is established by Víšek (2006). He considers the estimator to be nonlinear, since a subset of the data is considered, and establishes consistency using two different approaches: (1) using an asymptotic linearity argument and (2) using the uniform law of large numbers of Andrews (1987). The conditions given in Víšek (2006) are general; however, our approach establishes consistency for a much larger class of models and estimators.

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 can be found from Theorem 1 by taking $p = \infty$ and $q = 1$. These are given in the following corollary.

Corollary 3. *If*

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

C2: $E(|e|) < \infty$ *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 *that is symmetric about 0 and strictly decreasing on* $\mathbb{R}^+$,

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

We next compare the result in Corollary 3 with the one given by Oberhofer (1982). Oberhofer proves the *weak* consistency by imposing the following conditions.

O1: If $\boldsymbol{\Theta}^*$ is a closed set not containing $\boldsymbol{\theta}_0$, then there exist numbers $\epsilon > 0$ and n_0 such that for all $n \geq n_0$

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

for all such $\boldsymbol{\Theta}^*$ where $l_i(\boldsymbol{\theta}) = f(\mathbf{x}_i;\boldsymbol{\theta}) - f(\mathbf{x}_i;\boldsymbol{\theta}_0)$.

O2: $E(|e|) < \infty$ and $E([f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)]^2) < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and

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

Here O3 is weaker than C3. However, O2 is stronger than C2. Following similar contrapositive arguments as in the least squares case, we can easily show that O1 is also stronger

than C1 (see also Oberhofer (1982) p. 318). For a detailed discussion of this and sufficient conditions for O1, the reader is referred to Oberhofer (1982).

Signed-Rank Wilcoxon

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 3 and thus omitted.

Corollary 4. *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* $r \geq 1$, $E(|e|^r) < \infty$ *and* $E(|f(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x};\boldsymbol{\theta}_0)|^r) < \infty$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *and*

D3: G *has a density* g *that is symmetric about 0 and strictly decreasing on* $\mathbb{R}^+$,

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

Remark 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. These scores were first proposed by Fraser (1957). Since* φ^+ *needs to be bounded for our approach to work, our results do not directly extend to the case of normal scores. However, we may use Winsorized normal scores such as*

$$\varphi^+(u) = \begin{cases} \Phi^{-1}(-k), & \text{if} \quad u < -2k-1; \\ \Phi^{-1}(\frac{u+1}{2}), & \text{if} \quad -2k-1 \leq u < 2k-1; \\ \Phi^{-1}(k), & \text{if} \quad u \geq 2k-1\ . \end{cases}$$

Usually we take $k = 4$.

2.4 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. Assuming that the true value of the parameter to be estimated is in the interior of the parameter space $\boldsymbol{\Theta}$, breakdown represents a severe form of *inconsistency* in that the estimator converges to a point on the boundary of $\boldsymbol{\Theta}$ instead of $\boldsymbol{\theta}_0$.

Recall that $\mathbf{V} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \subset \mathbb{V}$ denotes the sample data points. Let $\mathcal{V}^m$ be the set of all data sets obtained by replacing any m points in $\mathbf{V}$ by arbitrary points. The finite sample breakdown point of an estimator $\widehat{\boldsymbol{\theta}}$ is defined as [see Donoho and Huber (1983)]

$$\varepsilon_n^*(\widehat{\boldsymbol{\theta}}, \mathbf{V}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathbf{Z} \in \mathcal{V}^m} |\widehat{\boldsymbol{\theta}}(\mathbf{Z}) - \widehat{\boldsymbol{\theta}}(\mathbf{V})| = \infty \right\}, \tag{2.11}$$

where $\widehat{\boldsymbol{\theta}}(\mathbf{V})$ is the estimate obtained based on the sample $\mathbf{V}$. In nonlinear regression, however, this definition of the breakdown point fails since ε^* 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, ε_+ and ε_-, which depend on the regression model, f. For any $\mathbf{x}_0 \in \mathbb{X}$, the upper and lower breakdown points are defined as

$$\varepsilon_+(f, \widehat{\boldsymbol{\theta}}, \mathbf{V}, \mathbf{x}_0) = \begin{cases} \min_{0 \le m \le n} & \left\{ \frac{m}{n} : \sup_{\mathbf{Z} \in \mathcal{V}^m} f(\mathbf{x}_0, \widehat{\boldsymbol{\theta}}(\mathbf{Z})) = \sup_{\boldsymbol{\theta}} f(\mathbf{x}_0, \boldsymbol{\theta}) \right\} \\ & \text{if } \sup_{\boldsymbol{\theta}} f(\mathbf{x}_0, \boldsymbol{\theta}) > f(\mathbf{x}_0, \widehat{\boldsymbol{\theta}}), \\ 1 & \text{otherwise,} \end{cases} \tag{2.12}$$

and

$$\varepsilon_-(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0)=\begin{cases}\min_{0\le m\le n} & \left\{\frac{m}{n}:\inf_{\mathbf{Z}\in\mathcal{V}^m} f(\mathbf{x}_0,\widehat{\boldsymbol{\theta}}(\mathbf{Z}))=\inf_{\boldsymbol{\theta}} f(\mathbf{x}_0,\boldsymbol{\theta})\right\}\\ & \text{if } \inf_{\boldsymbol{\theta}} f(\mathbf{x}_0,\boldsymbol{\theta})<f(\mathbf{x}_0,\widehat{\boldsymbol{\theta}}),\\ 1 & \text{otherwise .}\end{cases} \tag{2.13}$$

Let

$$\varepsilon(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0)=\min\{\varepsilon_+(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0),\varepsilon_-(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0)\}.$$

The finite sample breakdown point is now defined as

$$\varepsilon(f,\widehat{\boldsymbol{\theta}},\mathbf{V})=\inf_{\mathbf{x}_0\in\mathbb{X}}\left\{\varepsilon(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0)\right\}. \tag{2.14}$$

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

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

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

where $k\ge[n/2]+1$. Here $[b]$ stands for the greatest integer less than or equal to b. 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

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

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 straightforward. The proof is obtained by replacing $\text{med}_{1\leq i\leq n}$ with $n^{-1}\sum_{i=1}^{k}$ and m with $n-k$ in Stromberg and Ruppert's (1992) proof of Theorem 3. In the following, $\#(A)$ denotes the cardinality of the set A.

Theorem 2. *Assume for some fixed* $\mathbf{x}_0$ *there exist* $\tau_k \subset \{i : 1 \leq i \leq n\}$ *where* $\#(\tau_k) = 2n - [n/2] - k$ *such that*

$$\lim_{M\uparrow\infty}\Big\{\inf_{\{\boldsymbol{\theta}:f(\mathbf{x},\boldsymbol{\theta})>M\}}\{\inf_{i\in\tau_k} f(\mathbf{x}_i,\boldsymbol{\theta})\}\Big\} = \sup_{\boldsymbol{\theta}} f(\mathbf{x}_0,\boldsymbol{\theta})$$

Then

$$\varepsilon_+(f,\widehat{\boldsymbol{\theta}},\mathbf{V},\mathbf{x}_0) \geq \frac{n-k+1}{n}.$$

Theorem 2 establishes that even when the regression function f lies on the boundary for a portion of the data, the bias of the estimator of $\boldsymbol{\theta}_0$ remains within reasonable bounds if trimming is implemented. The following corollary gives the asymptotic (as $n \to \infty$) breakdown point of $\widehat{\boldsymbol{\theta}}_n$.

Corollary 5. *Let* $\alpha = \sup\{u : \varphi^+(u) > 0\}$. *The asymptotic breakdown point of* $\widehat{\boldsymbol{\theta}}_n$ *is at least* $1-\alpha$.

This is reminiscent of the breakdown point of a linear function of order statistics which is equal to the smaller one of the two fractions of mass at either ends of the distribution which receive weights equal to zero (Hampel, 1971). The same result obtained in Corollary 5 was given by Hampel (1971) for one-sample location estimators based on linear functions of order statistics (see sec. 7 (i) of Hampel (1971)).

Consider the class of models with the form $f(x,\boldsymbol{\theta}) = g(\beta_0 + \beta_1 x)$, where $(\beta_0,\beta_1) \in \mathbb{R}^2$ and $g(t)$ is monotone increasing in t. This class of models is considered by Stromberg and Ruppert (1992) and contains popular models like the logistic regression model $g(\beta_0,\beta_1 x) = \{1+\exp(-(\beta_0+\beta_1 x))\}^{-1}$. A breakdown point of $1-\alpha$ can be achieved if $\widehat{\boldsymbol{\theta}}_n$ is obtained via a minimization of (2.2) with $a_n(i) = \varphi^+(i/(n+1))$ such that $\alpha = \sup\{u : \varphi^+(u) > 0\}$.

Remark 5. *A definition of breakdown based on 'badness measures' which includes the definition given by Stromberg and Ruppert (1992) was given by Sakata and White (1995). Under our assumptions this definition reduces to the one used in the current paper as shown in Theorem 2.3 of Sakata and White (1995).*

Chapter 3

Bounded Influence Nonlinear Signed-Rank Regression

3.1 Introduction

As in the previous chapter, let us consider the following nonlinear regression model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + e_i, \qquad 1 \leq i \leq n \,, \tag{3.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is a vector of parameters, $\mathbf{x}_i$ is a vector of independent variables in a vector space $\mathbb{X}$, and f is a real-valued function defined on $\mathbb{X} \times \boldsymbol{\Theta}$. Let $\mathbf{V} = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ be the set of sample data points. Note that $\mathbf{V} \subset \mathbb{V} \equiv \mathbb{R} \times \mathbb{X}$. 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 a twice continuously differentiable function of $\boldsymbol{\theta}$ for each $\mathbf{x} \in \mathbb{X}$ and a measurable function of $\mathbf{x}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The errors e_i are assumed to be iid with a distribution function G.

The asymptotic normality of the least squares (LS) estimator of $\boldsymbol{\theta}_0$ has been discussed in Jennrich (1969), Wu (1981), and Wang (1996) among others. The asymptotic normality of the least absolute deviations (LAD) estimator of $\boldsymbol{\theta}_0$ is discussed in Wang (1995). However, as pointed out in Haupt and Oberhofer (2009), the treatment of Wang (1995) and Wang (1996) were missing some necessary global conditions. The estimator that will be introduced in this chapter is based on a generalized form of the signed-rank objective function. It provides a unified treatment of a class of estimators including those considered in Wang (1995) and Wang (1996). Moreover, we show how a weight functions can be incorporated to obtain estimators with bounded influence function (Hampel, 1974). Simply stated, the influence function represents the amount of change in the estimator caused by infinitesimal

contamination in the data. Thus it is a measure of the sensitivity of an estimator to outliers and it is desired that this function be bounded.

Rank-based estimators of linear models (where $f(\mathbf{x}, \boldsymbol{\theta}_0) = \mathbf{x}'\boldsymbol{\theta}_0$ in (3.1)) have been studied extensively. Jaeckel (1972) gave a general class of rank estimators for linear regression parameters that are efficient and robust to outliers in the response space. These include the Wilcoxon estimator which is equal to the median of pairwise slopes $(Y_j - Y_i)/(x_j - x_i)$ for the case of simple linear regression. These estimators, however, were found to be sensitive to outliers in the $\mathbf{x}$ direction (Hettmansperger et al., 2000; Hettmansperger and McKean, 1998); thus having an unbounded influence function. Sievers (1983) introduced weighted Wilcoxon estimators that were later shown to possess a bounded influence function by Naranjo and Hettmansperger (1994). Chang et al. (1999) provided one-step estimators that have high breakdown point based on the weighted Wilcoxon pseudonorm, where the weights depend on a robust and consistent estimator of $\boldsymbol{\theta}_0$.

The signed-rank (SR) estimator of the slope parameter in the linear model is also efficient and robust to outliers in the y direction but sensitive to outliers in the $\mathbf{x}$ direction (Hettmansperger et al., 2000). As the Wilcoxon estimator, the SR estimator is suitable when dealing with datasets from studies with controlled designs. However, it may be adversely affected when exploring datasets based on uncontrolled studies. To address the lack of robustness in the $\mathbf{x}$ direction, Tableman (1990) provided a one step signed-rank estimator for the linear model that has a bounded influence function. The results of Tableman (1990) were motivated by the work of Krasker and Welsch (1982) who gave a class of M-estimators with bounded influence function for linear regression estimation. A framework similar to Tableman (1990) has been investigated by Wiens and Zhou (1994) who provided bounded-influence rank estimators in the linear model using a general form of the SR objective function. They also show how the efficiency can be optimized by appropriate choices of scores and weights under a boundedness constraint on the influence function.

For the general nonlinear model given in (3.1), Abebe and McKean (2007) studied the asymptotic properties of the Wilcoxon estimator of $\boldsymbol{\theta}_0$. Just as in linear models, this estimator was shown to be efficient but sensitive to local changes in the direction of $\mathbf{x}$. Jurečková (2008) also studied the asymptotic properties of the rank estimator of $\boldsymbol{\theta}_0$ in (3.1). Her approach takes advantage of the asymptotic equivalence of regression quantiles and regression rank scores to provide rank scores based on the regression function. The approach results in a restricted set of scores. Also, the resulting estimator does not possess a bounded influence function.

In this chapter, we propose a class of rank-based estimators of $\boldsymbol{\theta}_0$ in (3.1) based on the minimization of a weighted signed-rank objective function. In contrast with the approach of Abebe and McKean (2007) and Jurečková (2008), this approach allows for a set of scores generated by any nondecreasing bounded score function that has at most a finite number of discontinuities. Also, by utilizing the theory of Sobolev spaces, this approach removes certain restrictive assumptions such as compactness of $\mathbb{X}$, Lipschitz continuity of the regression function, boundedness of the first derivative of the density of the error distribution that were needed in the work of Jurečková (2008). Our objective function is very general. For instance, the LS objective function is a special case of our objective function. However, the objective function of Jurečková (2008) does not include the LS objective function. We also show how Krasker-Welsch type weights (Krasker and Welsch, 1982) can be defined based on the regression function f to result in a bounded influence function. Moreover, simulation studies show that the proposed weighted estimators are also efficient attaining a relative efficiency of .955 versus least squares when G is Gaussian.

Other robust approaches to nonlinear regression include Stromberg (1993) who provided computational algorithms for computing high breakdown nonlinear regression parameters using the least median of squares (Rousseeuw, 1984) and MM (Yohai, 1987) estimators. Stromberg (1995) establishes the consistency of the least trimmed squares (LTS) estimator

for the nonlinear model in (3.1). The LTS was shown to have a high breakdown point by Stromberg and Ruppert (1992).

For linear models, the estimator proposed in this chapter can be regarded as a generalization of the objective function of Tableman (1990) to include other norms such as weighted LAD and LS. Moreover, since we do not restrict ourselves to the linear model, not only is it an extension of signed rank estimators for the linear model to the nonlinear regression case, but it is also a generalization of LAD and LS type estimators for the nonlinear regression model.

The remainder of the chapter is organized as follows. Our proposed estimator is given in Section 3.2. Section 3.2 also contains asymptotic and robustness results concerning the proposed weighted estimator. Section 3.3 gives the results using plug-in estimator of the weights based on a consistent estimator of the regression parameter. Real data and simulation examples are given in Section 3.4. Section 3.5 provides a discussion. Proofs and technical results are given in the appendix.

3.2 Weighted SR Estimator

Consider the signed-rank (SR) estimator, $\widehat{\boldsymbol{\theta}}_S$, of $\boldsymbol{\theta}_0$ in equation (3.1) that minimizes $T_n^+(\boldsymbol{\theta}) = \sum_{i=1}^n R_i|z_i(\boldsymbol{\theta})|$, where $z_i(\boldsymbol{\theta}) = y_i - f(\mathbf{x}_i, \boldsymbol{\theta})$ and $R_i = \#\{j : |z_j(\boldsymbol{\theta})| \leq |z_i(\boldsymbol{\theta})|\}$ is the rank of $|z_i(\boldsymbol{\theta})|$, $i = 1, \ldots, n$. The least squares (LS) and least absolute deviation (LAD) estimators of $\boldsymbol{\theta}_0$ minimize $\sum_{i=1}^n z_i^2(\boldsymbol{\theta})$ and $\sum_{i=1}^n |z_i(\boldsymbol{\theta})|$, respectively. It is well known that the LS estimator is sensitive to outliers in both x and y directions while the SR and LAD estimators are sensitive to outliers in the x-direction. There is clearly a need for a method that is not sensitive to outliers in both x and y directions. We obtain this by considering a weighted form of the SR estimator.

We define the weighted SR (WSR) estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ to be any vector $\boldsymbol{\theta}$ minimizing

$$D_n(\mathbf{V}, w, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, \boldsymbol{\theta}_0) a_n(i) \rho(|z(\boldsymbol{\theta})|_{(i)}) \tag{3.2}$$

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 $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is continuous and strictly increasing. The numbers $a_n(i)$ are scores generated as $a_n(i) = \varphi^+(i/(n+1))$, for some bounded and non-decreasing score function $\varphi^+ : (0,1) \to \mathbb{R}^+$ that has at most a finite number of discontinuities. The function $w : \mathbb{X} \times \boldsymbol{\Theta} \to \mathbb{R}^+$ is a continuous weight function. Because $D_n(\mathbf{V}, w, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, Lemma 2 of Jennrich (1969) implies the existence of a minimizer of $D_n(\mathbf{V}, w, \boldsymbol{\theta})$.

It is clear that weighted LS and LAD are special cases of WSR. Weighted LS is obtained by taking $\varphi^+ \equiv 1$ and $\rho(t) = t^2,\ t \geq 0$ while weighted LAD is obtained by taking $\varphi^+ \equiv 1$ and $\rho(t) = t$. In our analyses, however, LS and LAD refer to the unweighted versions obtained by taking $w \equiv 1$.

In the following, we will establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ and discuss how weights can be used to obtain a bounded influence function. As given in (3.2), the weights depend on the unknown true parameter $\boldsymbol{\theta}_0$. This will make our derivations cleaner. However, to be of practical use, the weights would have to be estimated. In Section 3.3, we will discuss a plug-in estimator of the weights based on a consistent estimator of $\boldsymbol{\theta}_0$ and how estimators based on these estimated weights have the same asymptotic properties as their counterparts based on 'true' weights.

3.2.1 Preliminaries

The following definitions and notations will be used throughout this paper. Let Ω be a domain. We denote by $L^p(\Omega, P)$, $1 \leq p \leq \infty$, the space of P-measurable functions on Ω for which $\int_\Omega |h|^p dP < \infty$ with the usual modification for $p = \infty$. $C^\infty(\Omega)$ is the space of smooth(infinitely differentiable) functions defined in Ω, $D(\Omega)$ is the space of smooth functions

with compact support in Ω and $L^1_{loc}(\Omega)$ is the space of locally integrable functions in Ω; that is, functions that are integrable in any compact subset of Ω. Let $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}_0^n$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, be a multi-index. The differential operator is defined as

$$D^{\alpha}_{\boldsymbol{\theta}} = \frac{\partial^{|\alpha|}}{\partial\theta_1^{\alpha_1}...\partial\theta_n^{\alpha_n}} \ ,$$

where $|\alpha| = \sum_{i=1}^{n} \alpha_i$ and $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$. Let $\Gamma \in L^1_{loc}(\Omega)$. Given $\alpha \in \mathbb{N}_0^n$, a function $\eta \in L^1_{loc}(\Omega)$ is called the α^{th}-weak derivative of Γ if for all $\psi \in D(\Omega)$

$$\int_{\Omega} \Gamma D^{\alpha}_{u} \psi du = (-1)^{|\alpha|} \int_{\Omega} \eta\psi du \ ,$$

and we put $\eta = D^{\alpha}_{\boldsymbol{\theta}}\Gamma$.

As an example, consider $\Gamma(u) = |u|$. Clearly, Γ is not differentiable in the usual sense at 0. But $\Gamma \in L^1_{loc}(\mathbb{R})$, Γ is weakly differentiable and $\Gamma^{'}(u) = sgn(u)$.

Let $m \in \mathbb{N}_0$ and $1 \leq p \leq \infty$. The Sobolev space denoted by $W^{m,p}(\Omega)$ is defined as

$$W^{m,p}(\Omega) = \{\Gamma \in L^p(\Omega) : \quad D^{\alpha}_{\boldsymbol{\theta}}\Gamma \in L^p(\Omega) \text{ with } |\alpha| \leq m\} \ .$$

Given a function $K \in L^1$ such that $\int_{\mathbb{R}^n} K(x)dx = 1$, let $K_\delta(x) = \delta^{-n}K(x/\delta)$. The family of functions $\{K_\delta, \ \delta > 0\}$, is called a mollifier with kernel K and K_δ is known as the Friedrichs' mollifier. Some important facts related to Sobolev spaces that may be useful in our discussion are listed below without proofs. A detailed discussion of these can be found in Brezis (1983) and Adams (1975).

(S_1) $K_\delta \in C^\infty(\mathbb{R}^n)$, $supp(K_\delta) = \{x \in \mathbb{R}^n : \ \|x\| \leq \delta\}$, $K_\delta \geq 0$, and $\int_{\mathbb{R}^n} K_\delta(x)dx = 1$. Here $supp(K_\delta)$ denotes the support of K_δ.

(S_2) (Regularization Theorem)

Let K_δ be a Friedrichs' mollifier. If $\Gamma \in L^1_{loc}(\mathbb{R}^n)$, then the convolution product

$$\Gamma * K_\delta(x) = \int_{\mathbb{R}^n} \Gamma(x-y)K_\delta(y)dy$$

exists for all $x \in \mathbb{R}^n$. Moreover, $\Gamma * K_\delta \in C^\infty(\mathbb{R}^n)$, $supp(\Gamma * K_\delta) \subset supp(\Gamma) + B'(0,\delta)$ where $B'(0,\delta) = \{x \in \mathbb{R}^n : \|x\| \leq \delta\}$, $D^\alpha(\Gamma * K_\delta) = D^\alpha K_\delta * \Gamma$ and $supp(D^\alpha K_\delta) \subset supp(K_\delta)$. Also if $\mathcal{M}$ is a compact set of points of continuity of Γ, then $\Gamma * K_\delta \to \Gamma$ uniformly on $\mathcal{M}$ as $\delta \to 0$.

(S_3) Let K_δ be a Friedrichs' mollifier. Let $\Gamma \in W^{m,p}(\Omega)$ for $1 \leq p \leq \infty$. Then, $\Gamma * K_\delta \to \Gamma$ in $L^p(\Omega)$ and $\Gamma * K_\delta \to \Gamma$ in $W^{m,p}(\omega)$ as $\delta \to 0$ for all $\omega \subset\subset \Omega$. $\omega \subset\subset \Omega$ means ω is open, the closure of ω, $\bar{\omega}$ is compact and $\bar{\omega} \subset \Omega$.

3.2.2 Consistency

Let $(\Omega', \mathcal{F}, P)$ be a probability space. For $i = 1, \ldots, n$, assume that $\mathbf{x}_i$ and $e_i = y_i - f(\mathbf{x}_i; \boldsymbol{\theta}_0)$ are independent random variables (carried by $(\Omega', \mathcal{F}, P)$) with distributions H and G, respectively. Setting $\tilde{G}_\theta$ to denote the distribution of $|z(\theta)|$, we can rewrite $D_n(\mathbf{V}, w, \boldsymbol{\theta})$ as

$$D_n(\mathbf{V}, w, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, \boldsymbol{\theta}_0)a_n(i)(\rho \circ \tilde{G}^{-1}_{\boldsymbol{\theta}})(\xi_{(i)})$$

where $\xi_{(i)}$ are order statistics from the uniform $U(0,1)$ distribution.

Theorem 3. *Let*

(I_1) $P(f(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}_0)) < 1$ *for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,

(I_2) $w \in L^p(\mathbb{X} \times \boldsymbol{\Theta})$ *and there exists a function* $h \in L^q(\mathbb{V})$ *such that* $|\rho(\tilde{G}^{-1}_\theta(v))| \leq h(v)$, *for all* $\theta \in \Theta$ *and all* $1 \leq p, q \leq \infty$ *such that* $1/p + 1/q = 1$, *and*

(I_3) G *has a density* g *that is symmetric about 0 and strictly decreasing on* $\mathbb{R}^+$.

Then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

Before giving the proof, we state the following lemma without proof. The proof of this Lemma may be constructed following Lemma 2.

Lemma 4. *Under assumptions* $(A_1)-(A_3)$, $D_n(\mathbf{V}, w, \boldsymbol{\theta}) \xrightarrow{a.s.} \mu(\boldsymbol{\theta})$ *a.e.* $\mathbb{V}$, *uniformly for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *where* $\mu : \boldsymbol{\Theta} \to \mathbb{R}$ *is a function satisfying* $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \mu(\boldsymbol{\theta}) > \mu(\boldsymbol{\theta}_0)$ *for any* $\boldsymbol{\Theta}^*$ *a closed subset of* $\boldsymbol{\Theta}$ *not containing* $\boldsymbol{\theta}_0$.

Proof. By Lemma 1 of Wu (1981), to establish the consistency of $\widehat{\boldsymbol{\theta}}_n$, it is sufficient to show that

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} \big(D_n(\mathbf{V}, w, \boldsymbol{\theta}) - D_n(\mathbf{V}, w, \boldsymbol{\theta}_0)\big) > 0 \quad \text{a.s.} \tag{3.3}$$

for any $\boldsymbol{\Theta}^*$ a closed subset of $\boldsymbol{\Theta}$ not containing $\boldsymbol{\theta}_0$. To that end let $A_n(\mathbf{V}, w, \boldsymbol{\theta}) = D_n(\mathbf{V}, w, \boldsymbol{\theta}) - \mu(\boldsymbol{\theta})$, $B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta}_0)$, and $C_n(\mathbf{V}, w, \boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}_0) - D_n(\mathbf{V}, w, \boldsymbol{\theta}_0)$.

By Lemma 2, we have $A_n(\mathbf{V}, w, \boldsymbol{\theta}) \xrightarrow{a.s.} 0$ uniformly for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$, $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} B(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > 0$, and $\liminf_{n\to\infty} C_n(\mathbf{V}, w, \boldsymbol{\theta}_0) = 0$ a.s. For the statement given in (3.3) to hold, it suffices to show that $\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^*} A_n(\mathbf{V}, w, \boldsymbol{\theta}) = 0$ a.s. $A_n(\mathbf{V}, w, \boldsymbol{\theta})$, being uniformly convergent and continuous on a compact set $\boldsymbol{\Theta}^*$, is equicontinuous on $\boldsymbol{\Theta}^*$ a.e. $\mathbb{V}$. This gives $\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}^*} \{A_n(\mathbf{V}, w, \boldsymbol{\theta}^*)\} = 0$ a.s. and the proof is complete. □

Assumption (I_1) is a very weak condition needed for $\boldsymbol{\theta}_0$ to be identified. The linear version of (I_1) was given by Hössjer (1994) as $P(|\boldsymbol{\theta}'\mathbf{x}| = 0) < 1$ for any $\boldsymbol{\theta} \neq \mathbf{0}$ under the assumption that $\boldsymbol{\theta}_0 = \mathbf{0}$. Since φ^+ is bounded, by (I_2), we have $\|w\varphi^+\|_p < \infty$. Moreover, (I_2) and Hölder's inequality ensure that the product $(w\varphi^+)(\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1})$ is integrable. (I_3) admits a wide variety of error distributions examples of which are the normal, double exponential and Cauchy distributions with location parameter equal to 0.

3.2.3 Asymptotic Normality

Write $\Gamma_{\boldsymbol{\theta}}(t) = \rho[\tilde{G}_{\boldsymbol{\theta}}^{-1}(t)]$ and $\lambda_i = w(\mathbf{x}_i, \boldsymbol{\theta}_0)a_n(R_{\xi_i})$ where R_{ξ_i}, $i = 1, ..., n$ are the rank of $\xi_1, ..., \xi_n$. Then (3.2) can be written as

$$D_n(\mathbf{V}, w, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, \boldsymbol{\theta}_0)a_n(i)(\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1})(\xi_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i \Gamma_{\boldsymbol{\theta}}(\xi_i) \ .$$

By (I_2), $\|\lambda_i\|_p < \infty$ for $1 \leq p \leq \infty$. Now set $\Psi_n(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}}^{\alpha} D_n(V, w, \boldsymbol{\theta})$ and $\phi_{\boldsymbol{\theta}}(t) = D_{\boldsymbol{\theta}}^{\alpha}\Gamma_{\boldsymbol{\theta}}(t)$ for $|\alpha| = 1$. Since the dependence of $\phi_{\boldsymbol{\theta}}$ on y is only through $z(\boldsymbol{\theta})$, we will suppress y in the notation and write $\phi_{\boldsymbol{\theta}}(\mathbf{x})$. Now denote the $n \times p$ matrix $\mathbf{X}^*$ by $\mathbf{X}^* = \left(\phi_{\boldsymbol{\theta}_0}(\mathbf{x}_1), ..., \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_n)\right)$ and define h_{ii}^n to be the i^{th} diagonal component of $\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}$. Now $\widehat{\boldsymbol{\theta}}_n$ is a zero of

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i \phi_{\boldsymbol{\theta}}(\xi_i). \tag{3.4}$$

Thus $\widehat{\boldsymbol{\theta}}_n$ can be seen as a weighted M-estimator with weights $\lambda_1, \ldots, \lambda_n$. So, under some conditions, the asymptotic theory of the weighted M-estimation can be applied.

In addition to (I_1) - (I_3), consider the following conditions:

(I_4) $\boldsymbol{\theta} \to \Gamma_{\boldsymbol{\theta}}(t)$ is a map in $W^{3,p}(B)$, where B is a neighborhood of $\boldsymbol{\theta}_0$ for every fixed t.

(I_5) There exist functions $\psi_\alpha \in W^{2,p}(\mathbb{V})$ such that $|D_{\boldsymbol{\theta}}^{\alpha}\phi_{\boldsymbol{\theta}}(t)| \leq \psi_\alpha(t)$ for every $\boldsymbol{\theta} \in B$ and $|\alpha| \leq 2$.

(I_6) $A_{\boldsymbol{\theta}_0} = E\left[w(\mathbf{x}, \boldsymbol{\theta}_0)\varphi^+(\xi)\left[D_{\boldsymbol{\theta}}^{\alpha}\phi_{\boldsymbol{\theta}}(\xi)\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$, where $\xi \sim U(0,1)$, is a positive definite matrix for $|\alpha| = 1$.

(I_7) $\lim\limits_{n\to\infty} \max\limits_{1\leq i\leq n} h_{ii}^n \to 0$

Example 1. *The assumptions above allow us to define certain types of hybrid estimators which may be constructed in the interest of efficiency and robustness. One such estimator is one that behaves like an LS estimator for small absolute residuals and like an LAD estimator*

for large absolute residuals. As an illustration, let us consider the one-dimensional case with $\boldsymbol{\Theta} = [a, b] \cup [b, c]$*, where* $a < b < c$ *are real numbers, and define*

$$\rho(|z(\theta)|) = \begin{cases} z^2(\theta), & \theta \in [a, b] \\ |z(\theta)| + z^2(b) - |z(b)|, & \theta \in [b, c] \ . \end{cases}$$

This function ρ *is strictly increasing and continuous as a function of* θ *but not differentiable in the usual sense at* $\theta = b$*. The test function* $\psi \in D(\boldsymbol{\Theta})$ *has to satisfy* $\psi(a) = \psi(b) = \psi(c) = 0$*. Using the definition of the weak derivative* γ *of* ρ *satisfies*

$$\int_{\boldsymbol{\Theta}} \rho(|z(\theta)|)\psi'(\theta)d\theta = -\int_{\boldsymbol{\Theta}} \gamma(\theta)\psi(\theta)d\theta \ .$$

But

$$\begin{aligned} \int_{\boldsymbol{\Theta}} \rho(|z(\theta)|)\psi'(\theta)d\theta &= \int_a^b z^2(\theta)\psi'(\theta)d\theta + \int_b^c |z(\theta)|\psi'(\theta)d\theta \\ &= -\left[\int_a^b 2z(\theta)\dot{f}(\theta, x)\psi(\theta)d\theta + \int_b^c sgn(z(\theta))\dot{f}(\theta, x)\psi(\theta)d\theta\right] \\ &= -\int_{\boldsymbol{\Theta}} \left[2z(\theta)\dot{f}(\theta, x)I_{[a,b]}(\theta) + sgn(z(\theta))\dot{f}(\theta, x)I_{[b,c]}(\theta)\right]\psi(\theta)d\theta \end{aligned}$$

where the second equality is from integration by parts and $\dot{f}(\theta, x) = \partial f(\theta, x)/\partial\theta$*. Thus*

$$\gamma(\theta) = 2z(\theta)\dot{f}(\theta, x)I_{[a,b]}(\theta) + sgn(z(\theta))\dot{f}(\theta, x)I_{[b,c]} \ .$$

Higher order derivatives can be computed similarly. In this case, (I_4) *and* (I_5) *are satisfied as long as* f *and its higher order derivatives (up to order 3) are bounded by integrable functions.*

Particularly, assumption (I_4) allows us to include a large variety of generalized functions that are not differentiable in the usual sense. The typical example is the absolute valued function given by $x \to |x|$ that includes the LAD and SR estimators, is not differentiable in the usual sence but is locally integrable, and therefore, weakly differentiable. It can be seen

that none of the computions above can be done without assumption (I_4). Assumption (I_5) ensures the integrability of $(w\varphi^+)$ $(D^\alpha_{\boldsymbol{\theta}}\phi_{\boldsymbol{\theta}})$ for any α such that $|\alpha| \leq 2$ for which the SLLN can be applied (see remark below). $A_{\boldsymbol{\theta}_0}$ in assumption (I_6) denotes the limiting Hessian matrix. (I_7) is a common assumption known as Noether's condition that ensures the asymptotic normality of $\Psi_n(\boldsymbol{\theta}_0)$.

Remark 6. *Under $(I_1), (I_2)$, and (I_3), Lemma 2 in the appendix gives the pointwise almost sure convergence of $D_n(\mathbf{V}, w, \boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If in addition (I_5) holds, then we have $[D^\alpha_{\boldsymbol{\theta}} D_n(\mathbf{V}, w, \boldsymbol{\theta})]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{a.s.} \mu_\alpha(\boldsymbol{\theta}_0)$ a.e. $\mathbb{V}$, where $\mu_\alpha(\boldsymbol{\theta}_0) \equiv E\left[w(\mathbf{x}, \boldsymbol{\theta}_0)\varphi^+(\xi)\left[D^\alpha_{\boldsymbol{\theta}}\phi_{\boldsymbol{\theta}}(\xi)\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]$ for any α such that $|\alpha| \leq 3$.*

The following theorem gives the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$. The approach of the proof is similar to that given in van der Vaart (1998) for M-estimators.

Theorem 4. *Under assumptions $(I_1) - (I_7)$,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0,\ A^{-1}_{\boldsymbol{\theta}_0}\Sigma_{\boldsymbol{\theta}_0}A^{-1}_{\boldsymbol{\theta}_0}),$$

where $\Sigma_{\boldsymbol{\theta}_0} = E\left[w(\mathbf{x}, \boldsymbol{\theta}_0)\varphi^+(\xi)\phi_{\boldsymbol{\theta}_0}(\xi)(\phi_{\boldsymbol{\theta}_0}(\xi))^T\right]$.

Proof. By (I_4), $\boldsymbol{\theta} \to \Gamma_\theta$ is a map in $W^{3,p}(B)$, properties (S_1) – (S_3) imply that $D(\bar{B})$ is dense in $W^{m,p}(B)$, where $\bar{B}$ is the closure of B. Thus, in the the following, we may assume without loss of generality that $\Gamma_{\boldsymbol{\theta}} \in D(\bar{B})$.

Implementing the Taylor expansion at $\boldsymbol{\theta}_0$ of $\Psi_n(\boldsymbol{\theta})$, we get

$$0 = \Psi_n(\widehat{\boldsymbol{\theta}}_n) = \Psi_n(\boldsymbol{\theta}_0) + \dot{\Psi}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)'\ddot{\Psi}_n(\gamma_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

where γ_n is a point between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$, $\dot{\Psi}_n(\boldsymbol{\theta}_0) = \left[D^\alpha_{\boldsymbol{\theta}}\Psi_n(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ for $|\alpha| = 1$ and $\ddot{\Psi}_n(\boldsymbol{\theta}_0) = \left[D^\alpha_{\boldsymbol{\theta}}\Psi_n(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ for $|\alpha| = 2$.

Now using (I_7) and by an application of the Cramér-Wold device (Serfling, 1980) and the central limit theorem for linear combinations of functions of order statistics (Hájek and

Šidák, 1967), $\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)$ converges to a multivariate normal distribution with mean 0 and covariance matrix $\Sigma_{\boldsymbol{\theta}_0}$. Also $\dot{\Psi}_n(\boldsymbol{\theta}_0)$ converges almost surely to $A_{\boldsymbol{\theta}_0}$ (see Remark 6) and hence in probability. By (I_4) and Theorem 3, $\lim_{n\to\infty} P(\{\gamma_n \in B\}) = 1$. So under the event $\{\gamma_n \in B\}$,

$$\|\ddot{\Psi}_n(\gamma_n)\| \leq C\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, \boldsymbol{\theta}_0)\psi(\xi_i)$$

where C stands for the bound of the score function. The right hand side of the above inequality is bounded in probability by the law of large numbers for n sufficiently large. These and the consistency of $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}_0$ give

$$-\Psi_n(\boldsymbol{\theta}_0) = \big[A_{\boldsymbol{\theta}_0} + o_p(1) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)O_p(1)\big](\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \big(A_{\boldsymbol{\theta}_0} + o_p(1)\big)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

since $(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)O_p(1) = o_p(1)O_p(1) \to 0$ in probability. Also with probability tending to 1, the matrix $A_{\boldsymbol{\theta}_0} + o_p(1)$ is invertible. Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\sqrt{n}\big(A_{\boldsymbol{\theta}_0} + o_p(1)\big)^{-1}\Psi_n(\boldsymbol{\theta}_0) = -\sqrt{n}A_{\boldsymbol{\theta}_0}^{-1}\Psi_n(\boldsymbol{\theta}_0) + o_p(1)\ . \tag{3.5}$$

The proof is complete by an application of Slutsky's lemma and noting that $\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)$ is asymptotically normal. □

In general, for estimators with $w \equiv 1$ and $\rho(t) = t$, we can simplify the asymptotic covariance matrix of $\sqrt{n}\widehat{\boldsymbol{\theta}}_n$ to obtain $\tau^2_{\varphi^+}\left\{E[\nabla f(\mathbf{x};\boldsymbol{\theta}_0)\nabla f(\mathbf{x};\boldsymbol{\theta}_0)^T]\right\}^{-1}$, where

$$\tau^2_{\varphi^+} = \frac{\int_0^1\{\varphi^+(u)\}^2du}{\left(\int_0^1 \varphi^+(u)\varphi_g^+(u)du\right)^2}\ ,$$

with $\varphi_g^+(u) = -g'(G^{-1}(\frac{u+1}{2}))/g(G^{-1}(\frac{u+1}{2}))$. It is easy to show that $\tau^2_{\varphi^+} = \{2g(0)\}^2$ for the LAD estimator and $\tau^2_{\varphi^+} = \left(\sqrt{12}\int_{-\infty}^{\infty}\{g(t)\}^2dt\right)^{-2}$ for the SR estimator. The asymptotic covariance matrix of the LS estimator is $\sigma^2\left\{E[\nabla f(\mathbf{x};\boldsymbol{\theta}_0)\nabla f(\mathbf{x};\boldsymbol{\theta}_0)^T]\right\}^{-1}$, where $\sigma^2 = \int u^2dG$ is the error variance.

For $\boldsymbol{\Theta} \subset \mathbb{R}^p$, two estimators can be compared using the asymptotic relative efficiency (ARE), which is the reciprocal of ratio of their asymptotic covariance matrices to the power $1/p$ (Serfling, 1980). The evaluation of the ARE is very complicated in general. We consider a simulation study with G taken to be the t distribution in Section 3.4. The ARE may be determined in closed form for comparing the simpler estimators such as SR versus LS or LAD for some error distributions such as the Cauchy, logistic, and normal. These are the same as those for comparing the signed-rank, the sign, and the t test in the location problem and are studied extensively in Hettmansperger and McKean (1998).

3.2.4 Robustness

In this section, we discuss conditions needed for the boundedness of the influence function. If we can write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda(\boldsymbol{\theta}_0; y_i, \mathbf{x}_i) + o_p(1) \ ,$$

then the influence function of $\widehat{\boldsymbol{\theta}}_n$ at a given point $(y_0, \mathbf{x}_0)$ is $\Lambda(\boldsymbol{\theta}_0; y_0, \mathbf{x}_0)$ (e.g. Corollary 3.5.7 of Hettmansperger and McKean (1998)). However, by equation (3.5) in the appendix we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-\lambda_i) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) + o_p(1) \ .$$

Thus $\Lambda(\boldsymbol{\theta}_0; y_0, \mathbf{x}_0) = -\lambda(\mathbf{x}_0, y_0) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_0)$, where $\lambda(\mathbf{x}_0, y_0) = w(\mathbf{x}_0)\varphi^+(G(z(\boldsymbol{\theta}_0))$. Assume that

(I_8) : there exists a constant $M > 0$ such that $\|w(\mathbf{x})\phi_{\boldsymbol{\theta}_0}(\mathbf{x})\| \leq M$.

Note that from the boundedness of the score function φ^+, for (I_8) to hold, it suffices that $\|w(\mathbf{x})\phi_{\boldsymbol{\theta}_0}(\mathbf{x})\|$ be bounded as a function of $\mathbf{x}$. This is the same condition as Assumption 8 of Coakley and Hettmansperger (1993) for $\phi_{\boldsymbol{\theta}_0}(\mathbf{x}) = \mathbf{x}$ corresponding to $f(\mathbf{x}, \boldsymbol{\theta}_0) = \mathbf{x}'\boldsymbol{\theta}_0$. Thus,

the choice of weights discussed in Section 6.1 of Coakley and Hettmansperger (1993) may be extended to our case.

Theorem 5. *Under* $(I_1) - (I_8)$, $\widehat{\boldsymbol{\theta}}_n$ *has a bounded influence function.*

3.3 Weight Specification

Based on the work of Giltinan et al. (1986) and Simpson et al. (1992) for the linear model, one may set

$$w(\mathbf{x}, \boldsymbol{\theta}_0) = \min\left[1, \frac{\eta}{d(\mathbf{x}, \boldsymbol{\theta}_0)}\right] \tag{3.6}$$

where $d(\mathbf{x}, \boldsymbol{\theta}_0) = (\mathbf{x}^* - \mathbf{m}_{\mathbf{x}^*})^T \mathbf{C}_{\mathbf{x}^*}^{-1} (\mathbf{x}^* - \mathbf{m}_{\mathbf{x}^*})$ is a robust Mahalanobis distance, with $\mathbf{m}_{\mathbf{x}^*}$ and $\mathbf{C}_{\mathbf{x}^*}$ being robust estimates of location and covariance of $\mathbf{x}^* = \phi_{\boldsymbol{\theta}_0}(\mathbf{x})$, respectively and η being some positive constant. As it can be seen from its definition, the weight function depends on the true parameter $\boldsymbol{\theta}_0$. So, in our study, we will consider plug-in estimators of $d(\mathbf{x}, \boldsymbol{\theta}_0)$ based on strongly consistent estimators of $\boldsymbol{\theta}_0$. As Coakley and Hettmansperger (1993) we will take resistant distances for $d(\mathbf{x}, \boldsymbol{\theta}_0)$ that use location and scale estimates based on the minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE) of Rousseeuw (1984) and Rousseeuw (1985), respectively. A discussion of these and other high breakdown estimators in the multivariate case is given in Hubert et al. (2008).

Analogous to Tableman (1990) and Krasker and Welsch (1982), one may use the weight function $w(\mathbf{x}, \boldsymbol{\theta}_0)$ in (4.13) by setting

$$d(\mathbf{x}, \boldsymbol{\theta}_0) = \mathbf{x}^{*T} B_{\boldsymbol{\theta}_0}^{-1} \mathbf{x}^*, \quad with \quad B_{\boldsymbol{\theta}_0} = E_{x,y}\left[w^2(\mathbf{x}, \boldsymbol{\theta}_0)\left(\varphi^+[2G(z(\boldsymbol{\theta}_0)) - 1]\right)^2 \mathbf{x}^{*T}\mathbf{x}^*\right]$$

$\mathbf{x}^* = \phi_{\boldsymbol{\theta}_0}(\mathbf{x})$ and $\eta = M/\zeta$ for ζ such that $\sup_t[\varphi^+(t)] = \zeta$. The difficulty in using this weight function for the nonlinear regression model is on estimating $d(\mathbf{x}, \boldsymbol{\theta}_0)$ for which the expression of $B_{\boldsymbol{\theta}_0}$ depends on the weight function $w(\mathbf{x}, \boldsymbol{\theta}_0)$. So we face a situation where we have to solve an implicit value problem. For linear models, however, since $d(\mathbf{x}, \boldsymbol{\theta}_0)$ can be specified

free of model parameters (see for example Section 6 of Coakley and Hettmansperger, 1993). An iterative scheme for computing $d(\mathbf{x}, \boldsymbol{\theta}_0)$ can be found in Krasker and Welsch (1982) and in Section 2 of Tableman (1990), and, under some suitable extra conditions (like the invertibility of the score function φ), one may extend the iterative computation of $d(\mathbf{x}, \boldsymbol{\theta}_0)$ to the nonlinear case. Another weight function is that considered by Wiens and Zhou (1994) and uses $d(\mathbf{x}_i, \boldsymbol{\theta}_0)$ that is equivalent to $\|A_{\boldsymbol{\theta}_0}^{-1}\mathbf{x}_i^*\|^{-1}$, where $\|\cdot\|$ is the Euclidean norm.

3.3.1 Plug-in Estimator of the Weight

Let $\tilde{\boldsymbol{\theta}}_n$ be the minimizer of $D_n(\mathbf{V}, w, \boldsymbol{\theta})$ with $w \equiv 1$; that is, the unweighted estimator of $\boldsymbol{\theta}_0$. $\widehat{\boldsymbol{\theta}}_S$ given at the beginning of Section 3.2 is one such estimator. Under $(A_1) - (A_7)$, $\tilde{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $n \to \infty$. Hence $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o(1)$ w.p. 1. Denote by $\tilde{w} = w(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}_n)$, $\tilde{\lambda}_i = w(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}_n)\varphi^+\Big(\frac{R_i}{n+1}\Big)$ and $\tilde{\Psi}_n(\boldsymbol{\theta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tilde{\lambda}_i \phi_{\boldsymbol{\theta}}(\xi_i)$ where $\tilde{\lambda}(\mathbf{x}_0, y_0) = w(\mathbf{x}_0, \tilde{\boldsymbol{\theta}})\varphi^+\Big(G(z(\boldsymbol{\theta}_0))\Big)$. Set $\widehat{\widehat{\boldsymbol{\theta}}}_n = \text{Argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\, \tilde{D}_n(\mathbf{V}, \tilde{w}, \boldsymbol{\theta})$, where $\tilde{D}_n(\mathbf{V}, \tilde{w}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) a_n(i)(\rho \circ \tilde{G}_{\boldsymbol{\theta}}^{-1})(\xi_{(i)})$. The following theorem shows that $\widehat{\widehat{\boldsymbol{\theta}}}_n$ and $\widehat{\boldsymbol{\theta}}_n$ have the same asymptotic properties.

Theorem 6. *Under* (I_2), $\tilde{\Psi}_n(\boldsymbol{\theta}_0) = \Psi_n(\boldsymbol{\theta}_0) + o(1)$ *w.p.1.*

Proof. From the fact that $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o(1)$ w.p.1 and by the continuity of the weight function w, we have $w(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) = w(\mathbf{x}_i, \boldsymbol{\theta}_0) + o(1)$ w.p.1. Now using this fact in the expression of $\tilde{\Psi}_n(\boldsymbol{\theta}_0)$, we get

$$\tilde{\Psi}_n(\boldsymbol{\theta}_0) = \Psi_n(\boldsymbol{\theta}_0) + o(1) \times \frac{1}{n}\sum_{i=1}^{n} \varphi^+\big(\frac{R_i}{n+1}\big)\phi_{\boldsymbol{\theta}_0}(\xi_i)\ .$$

Set $\Delta_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} \varphi^+\big(\frac{R_i}{n+1}\big)\phi_{\boldsymbol{\theta}_0}(\xi_i)$. Then, under (A_2), by the SLLN for functions of order statistics (Van Zwet, 1980), we have

$$\Delta_n(\boldsymbol{\theta}_0) \xrightarrow{a.s.} E[\varphi^+(\xi)\phi_{\boldsymbol{\theta}_0}(\xi)] < \infty\ ,$$

where $\xi \sim U(0,1)$. Thus $\Delta_n(\boldsymbol{\theta}_0)$ is bounded in probability and hence $o(1) \times \Delta_n(\boldsymbol{\theta}_0) \to 0$ w.p.1 as $n \to \infty$. Therefore $\tilde{\Psi}_n(\boldsymbol{\theta}_0) = \Psi_n(\boldsymbol{\theta}_0) + o(1)$ w.p.1. □

Theorem 7. *Under* $(I_1) - (I_7)$,

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \widehat{\boldsymbol{\theta}}_n) = o_p(1) \ .$$

Proof. As in the proof of Theorem 4, the Taylor expansion of $\tilde{\Psi}_n(\widehat{\widehat{\boldsymbol{\theta}}}_n)$ about $\boldsymbol{\theta}_0$ gives

$$0 = \tilde{\Psi}_n(\widehat{\widehat{\boldsymbol{\theta}}}_n) = \tilde{\Psi}_n(\boldsymbol{\theta}_0) + \dot{\tilde{\Psi}}_n(\boldsymbol{\theta}_0)(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0)' \ddot{\tilde{\Psi}}_n(\beta_n)(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) \ ,$$

where β_n lies on the line segment joining $\widehat{\widehat{\boldsymbol{\theta}}}_n$ and $\boldsymbol{\theta}_0$. Also under the assumptions of Theorem 6, one can show that $\dot{\tilde{\Psi}}_n(\boldsymbol{\theta}_0) = \dot{\Psi}_n(\boldsymbol{\theta}_0) + o(1)$ w.p.1. Now the fact that $w(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) = w(\mathbf{x}_i, \boldsymbol{\theta}_0) + o(1)$ w.p.1 and the SLLN of functions of the order statistics, we have $\dot{\tilde{\Psi}}_n(\boldsymbol{\theta}_0) \xrightarrow{a.s.} A_{\boldsymbol{\theta}_0}$. Note that under $(A_1) - (A_3)$, $\widehat{\widehat{\boldsymbol{\theta}}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$. From this and assumption (I_4), $\lim_{n\to\infty} P(\{\beta_n \in B\}) = 1$. So under the event $\{\beta_n \in B\}$, $\ddot{\tilde{\Psi}}_n(\beta_n)$ is bounded in probability. Hence

$$\begin{aligned}
\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) &= -\sqrt{n} A_{\boldsymbol{\theta}_0}^{-1} \tilde{\Psi}_n(\boldsymbol{\theta}_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-\tilde{\lambda}_i) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-\lambda_i) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) \ + \ o(1) \times A_{\boldsymbol{\theta}_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) + o_p(1) \ ,
\end{aligned}$$

where the last equality follows from $\tilde{\lambda}_i = \lambda_i + o(1)$ w.p.1 by the continuity of the weight function and the boundedness of φ^+. Clearly, by the central limit theorem, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i)$ converges to a normal distribution, thus bounded in probability. Therefore

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-\lambda_i) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) + o_p(1). \tag{3.7}$$

Now combining (3.5) and (3.7) yield $\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \widehat{\boldsymbol{\theta}}_n) = o_p(1)$.

□

Remark 7. *Theorem 4 and Theorem 7 show that* $\widehat{\widehat{\boldsymbol{\theta}}}_n$ *has the asymptotic Gaussian distribution given in Theorem 4. Moreover, from equation 3.7 in the appendix gives*

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\theta}}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-\lambda_i) A_{\boldsymbol{\theta}_0}^{-1} \phi_{\boldsymbol{\theta}_0}(\mathbf{x}_i) + o_p(1) \ .$$

Thus, $\widehat{\widehat{\boldsymbol{\theta}}}_n$ *has the same influence function as* $\widehat{\boldsymbol{\theta}}_n$*.*

3.4 Examples

All analyses in this section were performed using the R software environment (R Development Core Team, 2009). Estimates were found by means of the "nonlinear minimization subject to box constraints" algorithm (Gay, 1983, 1984) implemented in the function `nlminb` in the "stats" package of R with a range of starting values. We use weight functions based on robust Mahalanobis distances. MCD estimates of location and covariance were found using the fast MCD algorithm of Rousseeuw and Van Driessen (1999). Both MCD and MVE are implemented in the R function `cov.rob` in the package "MASS". For simplicity, all WSR estimators considered in the following use $\varphi^+(u) = u$ and $\rho(t) = t$. In this case $\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_s$.

3.4.1 Monte Carlo Simulation

Exponential Regression

Consider the one-parameter exponential regression model

$$y_i = e^{\theta_0 x_i} + \varepsilon_i \ , \qquad i = 1, \ldots, 25 \ . \tag{3.8}$$

We took $x_1, \ldots, x_{25}$ to be iid $N(0, 1)$. We then performed $B = 1000$ iterations where in each iteration $\varepsilon_1, \ldots, \varepsilon_{25}$ were randomly generated from the $N(0, .01)$ distribution and $y_1, \ldots, y_{25}$ were computed as $y_i = e^{x_i} + \varepsilon_i$. That is, the true value of θ_0 was taken to be 1. We considered two cases: an outlier in the y-direction by adding 50 to y_{20} and an outlier in the x-direction

by adding 5 to x_{20}. In each of the B repetitions we estimated θ_0 (giving $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$) using LS, LAD, SR, and WSR. For WSR estimation, we used the weight function

$$w(x, \tilde{\theta}) = \min\left\{1, \frac{\chi^2_{.95}(1)}{d(x, \tilde{\theta})}\right\}$$

where $\chi^2_{.95}(1)$ is the 95% percentile point of the $\chi^2(1)$ distribution and $d(x, \tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \sigma^2(x_i e^{\tilde{\theta} x_i} - \mu)^2$. Here $\tilde{\theta}$ is the SR estimate of θ_0 whereas μ and σ^2 are the MCD center and variance, respectively, of $xe^{\tilde{\theta}x}$. As the estimated value of θ_0, we took $B^{-1}\sum \hat{\theta}^{(i)}$ and as the estimate of the MSE, we took $B^{-1}\sum(\hat{\theta}^{(i)} - \theta_0)^2$. The results are given in Table 3.1.

Table 3.1: Average Estimates(MSEs) of $\theta_0 = 1$ in 1000 simulated samples

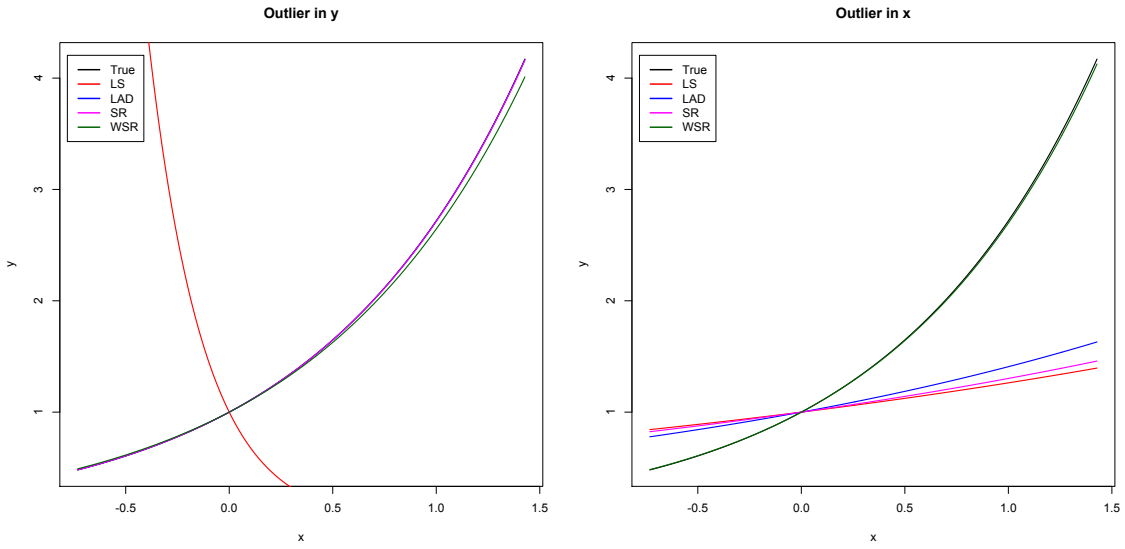
	LS	LAD	SR	WSR
y-outlier	-3.766(22.7108)	.999(.0001)	.999(.0001)	.973(.0031)
x-outlier	.233(.5878)	.342(.4327)	.264(.5413)	.993(.0001)

From Table 3.1, it is clear that the LS estimate is affected rather adversely by the presence of the y-outlier. We can also observe that the LS, LAD, and SR estimates are affected by the x-outlier while the WSR estimate remained close to $\theta_0 = 1$ in both cases. Figure 3.1 gives the plots of the fitted curves obtained using LS, LAD, SR, and WSR. It is clear from the plot that the WSR fit is not affected by either the outlying x or the outlying y values.

A Study of Relative Efficiency

It is well known that in linear models the relative efficiency (RE) of the signed rank estimator to the least squares estimator is $3/\pi \approx .955$ when G is the normal distribution. This RE increases with the heaviness of the tail of the error distribution. To study this for nonlinear models, we considered a simulation experiment involving the Michaelis-Menten and Gompertz functions. The Michaelis-Menten function describes the relation between velocity

Figure 3.1: Plot of fitted exponential curves including WSR



of reaction $f(x)$ of an enzyme with substrate concentration x. It is given by

$$f(x, \boldsymbol{\theta}) = \frac{\alpha x}{\beta + x} \ ,$$

where $\boldsymbol{\theta} = (\alpha, \beta)$, α is the maximum velocity of the reaction and β is the half-saturation constant. On the other hand the Gompertz model is defined by

$$f(x, \boldsymbol{\theta}) = \alpha \exp\{\mu e^{\beta x}\}$$

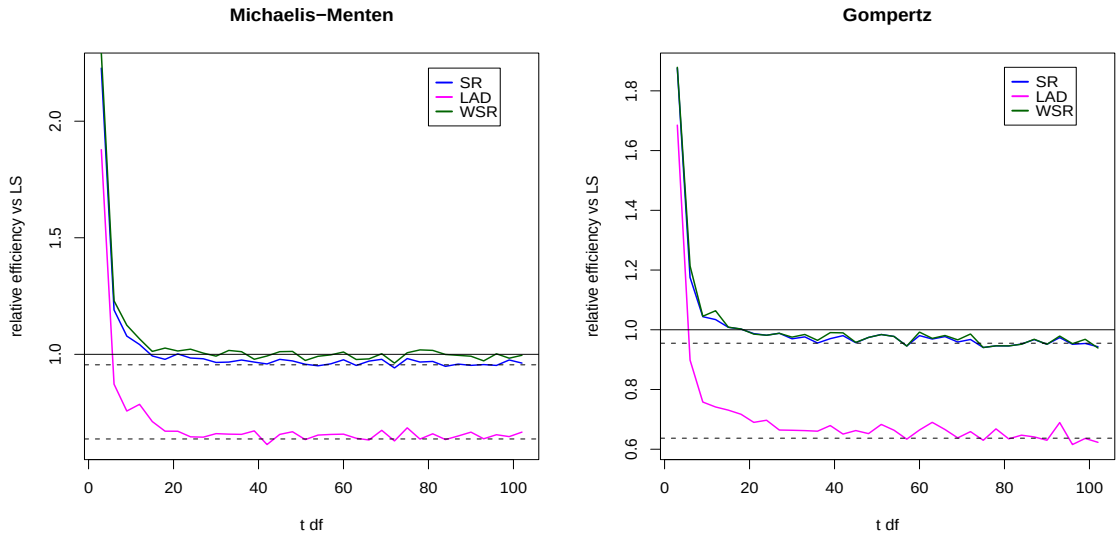
where $\boldsymbol{\theta} = (\alpha, \beta, \mu)$ is the vector of parameters of interest. One of the applications of this function is for modeling growth of tumors. In fact, tumors are cellular populations growing in a confined space where the availability of nutrients is limited. In this situation, $f(x)$ denotes the tumor size, α the carrying capacity, i.e., the maximum size that can be reached with the available nutrients, $\mu = \ln \frac{f(0)}{\alpha}$, where $f(0)$ is the tumor size at the starting observation time and β is the growth rate of the tumor.

For both models, $n = 100$ values of x were generated at random from the standard exponential distribution. Then for each of $B = 2000$ repetitions, n random errors ε were

generated from the t distribution with d degrees of freedom and n values of $y = f(x, \boldsymbol{\theta}) + \varepsilon$ were computed. This was done for degrees of freedom d ranging from 3 to 102 in steps of 3. The t distribution is ideal for simulating varying tail thicknesses as it provides distributions from the Cauchy ($d = 1$) to the normal ($d = \infty$). We are interested in estimating the parameter β in both models. The weighting scheme used for WSR estimation is the same as the one in the previous Monte Carlo experiment (MCD along with a $\chi^2_{.95}$ cutoff).

Figure 3.2 contains the plot of degrees of freedom versus the estimated relative efficiency given by ratios of estimated MSEs. The top dashed horizontal line is at $3/\pi$ and the bottom dashed horizontal line is at $2/\pi$. These values are the theoretical asymptotic relative efficiencies of SR versus LS and LAD versus LS, respectively, for the linear model. Estimators with relative efficiencies above the solid line are more efficient than the LS estimator. The plot shows that the SR and WSR estimators are more efficient than the LAD estimator. Moreover, we observe that the relative efficiencies approach the theoretical values under normality as the degree of freedom increases. We can also see that LAD, SR, and WSR estimators are more efficient than the LS estimator for heavy tailed error distributions. SR and WSR provide similar estimates of relative efficiency.

Figure 3.2: Relative Efficiency



3.4.2 Real Data

We now consider the Lakes Data given in Stromberg (1993). The data were collected from 29 lakes in Florida. Three variables were collected on each lake. The response is the mean annual nitrogen concentration (y). The predictors are the average influent nitrogen concentration (x_1) and the water retention time (x_2). The model recommended by the investigator is

$$y_i = \frac{x_{1i}}{1 + \alpha x_{2i}^{\beta}} + \varepsilon_i \ , \qquad i = 1, \ldots, 29 \ .$$

We fit the model using LS, LAD, SR, and WSR. For the weighted SR, we used the weight function

$$w(\mathbf{x}) = \min \left\{ 1, \frac{\chi^2_{.95}(2)}{d(\mathbf{x}, \tilde{\boldsymbol{\theta}})} \right\} \ .$$

We computed two versions of WSR. WSR1 uses weights where $d(\mathbf{x}, \tilde{\boldsymbol{\theta}})$ is taken to be the robust Mahalanobis distance constructed using the MVE estimates of the center and covariance of $\nabla_{\boldsymbol{\theta}} f(\mathbf{x}, \tilde{\boldsymbol{\theta}})$ with $f(\mathbf{x}, \boldsymbol{\theta}) = x_1/(1 + \alpha x_2^{\beta})$, $\boldsymbol{\theta} = (\alpha, \beta)$, and $\mathbf{x} = (x_1, x_2)$. WSR2 has the same setup as WSR1 except that it uses weights based on classical Mahalanobis distances. We take $\tilde{\boldsymbol{\theta}}$ to be the SR estimate of $\boldsymbol{\theta}$. The results obtained taking $\tilde{\boldsymbol{\theta}}$ as LS or LAD estimates are similar and hence not reported. Table 3.2 gives the estimates of the parameters, standard errors, and scale.

Table 3.2: Parameter estimates for Lakes Data

Method	Parameter	Estimate	SE	Z	p-value	Scale
LS	α	5.08423	1.93891	2.62220	0.00874	1.2637
	β	1.27852	0.35334	3.61835	0.00030	
LAD	α	3.88311	0.79337	4.89443	0.00000	1.0046
	β	1.29880	0.23082	5.62686	0.00000	
SR	α	5.09126	1.66082	3.06551	0.00217	1.3877
	β	1.22612	0.29373	4.17432	0.00003	
WSR	α	0.88588	0.58543	1.51320	0.13023	1.2904
	β	0.39018	0.17357	2.24799	0.02458	

For the original data, it is apparent from Table 3.2 that WSR1 gives estimates that are quite different from the other methods. The residual diagnostic plots given in Figure 3.3 can help us determine which of the results is the most appropriate. The figure contains plot of residuals from the fitted models versus x_1 (left panel) and residuals versus $\hat{y}$ (right panel). The plots for LAD and SR (not shown) are very similar to those of LS. We observe that all plots of residuals versus x_1 identify two potential outliers (denoted by solid squares) in the x_1 direction. However, these observations do not stand out as aberrant in the residual versus fits plot for LS whereas WSR2 identifies only one of them as a potential outlier. The WSR1 fit clearly identifies the two points as outliers. Stromberg (1993) also identifies the same points using the high breakdown least median of squares and MM estimators. The residual plot of WSR1 given in Figure 3.3 and that of MM given in Stromberg (1993) are quite similar. The "Outliers Removed" part of Table 3.2 gives the estimates and standard errors after removing the two outliers identified by WSR1. The estimates due to the different methods are quite similar to each other.

We removed the one outlier identified by WSR2 and refit all the models. LS, SR, and LAD fail to identify the remaining one outlier. However, WSR2 identifies the outlier. This indicates a potential masking problem associated with using the classical Mahalanobis distance in the weights. This is very much the same masking problem one faces when using hat matrix weights in linear regression. This issue is discussed in Wiens & Du (2000). In this case, the use of robust Mahalanobis distances appears to be a good solution to the masking problem.

3.5 Discussion

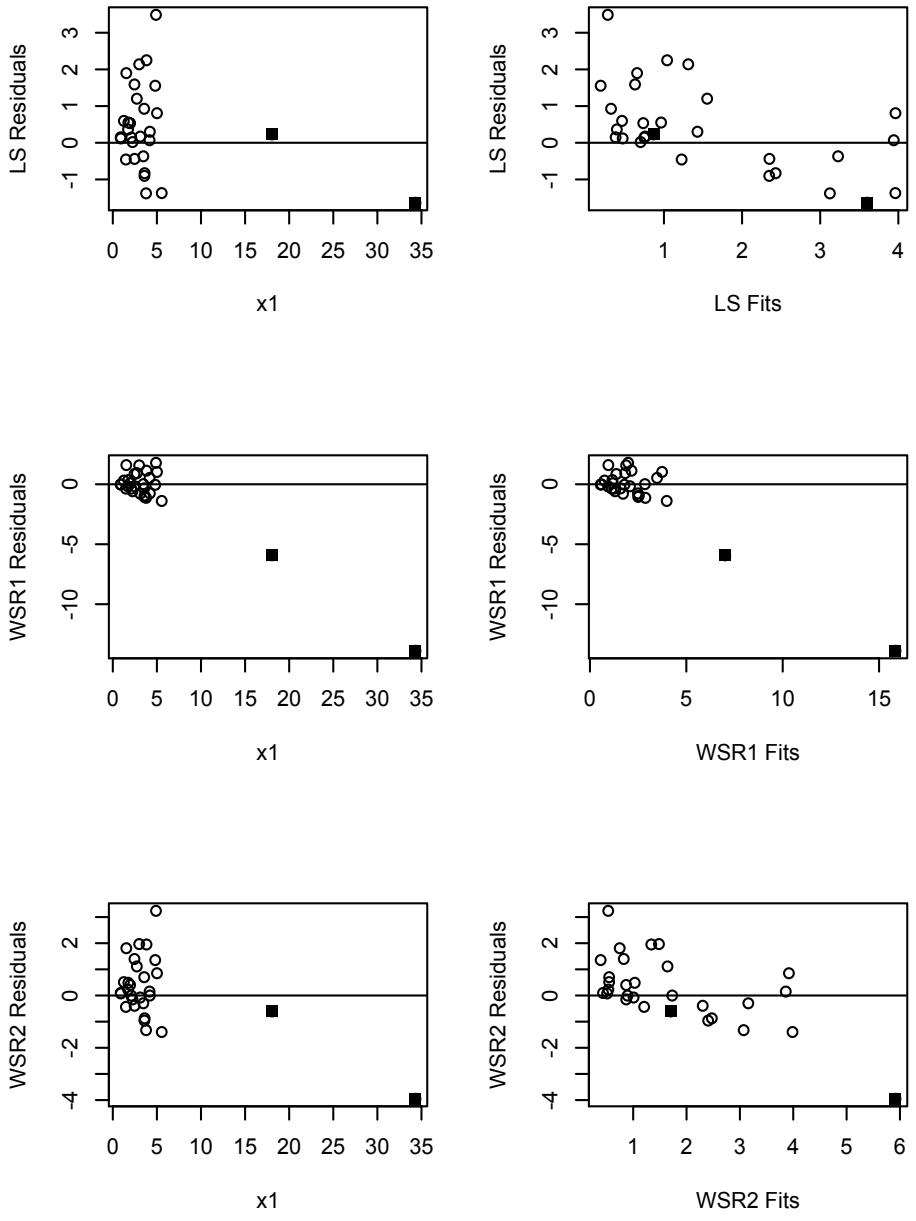
In this article, we proposed a rank-based analysis of nonlinear regression models. Our study uses a weighted generalized signed-rank dispersion function. The generalized signed-rank dispersion function results in a class of estimators including the signed-rank, least squares, and least absolute deviations estimators. These estimators do not have bounded

influence functions. The influence function of LAD and LS are unbounded in both response and design spaces while the that of the signed-rank is unbounded in design space. Thus, the signed-rank estimation procedure may not be suitable for studies with uncontrolled designs.

Added weights allow us to construct estimators with bounded influence. However, it is rather complicated to use the proposed weights directly since they cannot be expressed free of the regression parameter for nonlinear models. Our solution is to replace the regression parameter by its consistent estimator in the weight function. This results in estimators that are asymptotically equivalent to those in which the weights are based on the unknown true value of the parameter. The unweighted versions of our estimators are consistent; hence, they can be used to estimate the weights. We recommend weight functions that are based on robust Mahalanobis distances using robust estimates of location and covariance of the Jacobian matrix of the regression function. They are explicitly defined and are simple to construct. Moreover, as our simulation studies demonstrate, the weighted signed-rank estimators with weights based on robust Mahalanobis distances are efficient as well as robust. It is also shown that these weighted estimators can be useful in detecting outliers in nonlinear regression.

s

Figure 3.3: Residual plot versus x_1 and fitted values for Lakes data



Chapter 4

Rank Regression with Missing Response

4.1 Introduction

The problem of missing data is nowadays in the center of almost all statistical studies. It occurs in a wide array of application areas for several reasons. Among the reasons are:

- a sensor in a remote sensor network may be damaged and cease to transmit data
- certain regions of a gene microarray may fail to yield measurements of the underlying gene expressions due to scratches, finger prints, or manufacturing defects
- in a clinical trial, participants may drop out during the course of the study leading to missing observations at subsequent time points
- all applicable tests while diagnosing a patient may not be ordered by a doctor
- users of a recommender system rate extremely small fraction of available books, movies, or songs, leading to massive amount of missing data.

Also, data may be missing because equipment malfunctioned, the weather was terrible, or people got sick, or the data were not entered correctly. The type of missingness to be considered in this chapter is the missing response in the context of regression analysis that often arise in various experimental settings such as market research surveys, medical studies, opinion polls and socioeconomic investigations.

The statistical investigation of such a problem is a very difficult task since in most cases, missing data themselves contain either little or no information about the missing data mechanism (MDM). Fundamentally and most commonly used assumption about the

MDM is the MAR assumption discussed in Rubin (1976). The idea behind MAR is that the probability that a response variable is observed can depend only on the values of those other variables that have been observed. The scientific literature provides extensive studies and effective computational methods for handling missing data under the MAR assumption.

4.2 Model Definition

Consider the linear semi-parametric regression model

$$Y_i = X_i^\tau \beta + g(T_i) + \varepsilon_i, \qquad 1 \leq i \leq n \ , \tag{4.1}$$

where $\beta \in \mathcal{B}$ is a vector of parameters, X_i's are i.i.d p-variable random covariate vectors, T_i's are i.i.d univariable random covariates defined on $[0,1]$, the function $g : [0,1] \to \mathbb{R}$ is unknown and the model errors ε_i are independent with conditional mean zero given the covariates. Also, $0 < E(\varepsilon_i^2|Z_i) < \infty$ with $Z_i = (X_i, T_i)$. In this paper, we are interested in inference on the true value β_0 of the parameter β, when there are missing responses in the linear semi-parametric model (4.1).

This model has captured a lot of attention in recent years. An application of (4.1) to mouthwash experiment was given by Speckman (1988). An example using (4.1) is provided in Green and Silverman (1994). The least squares estimation approach of the setting discussed above was studied by Wang and Sun (2007). A semi-parametric mixed model for analyzing the CD4 cell count in HIV seroconverters was studied by Zeger and Diggle (1994). Model 4.1 has also been applied in several fields such as biometrics, see Gray (1994), econometrics and others. For complete data setting, this model has been extensively studied by many authors such as Heckman (1986), Speckman (1988), Robinson (1988), Rice (1986) among others. In the framework of model (4.1), Wang et al. (2004) developed inference tools in missing response case for the mean of Y based on the least squares estimation approach and under the MAR assumption. The historical method for constructing confidence interval

for the true mean of Y is the empirical likelihood method introduced by Owen (1990). As pointed out by many authors, this method has many advantages over others methods such as those based in normal approximations or the bootstrap (Hall and La Scala, 1990). Many authors have studied this framework including Kitamura (1997), Peng (2004), Wang et al. (2004), Xue and Zhu (2007b), Xue and Zhu (2007a), Xue and Zhu (2006), Wang and Rao (2002a), Sun et al. (2009), Chen and Hall (1993) among others.

When dealing with missing data, the main approach is to impute a plausible value for each missing datum and then analyze the results as if they were complete. In most of the regression problems, the commonly used approaches include linear regression imputation (Healy and Westmacott, 1956), nonparametric kernel regression imputation (Cheng, 1994; Wang and Rao, 2002b), semi-parametric regression imputation (Wang and Sun, 2007), among others. Wang and Sun (2007) also considered the semi-parametric regression imputation approach to estimate the true mean of Y.

The other well known approach for handling missing data is the inverse probability weighting. This approach has gained considerable attention as a way to deal with missing data problems. For a discussion of this approach, see Wang et al. (1997), Robins et al. (1994), Wang et al. (2004), Zhao et al. (1996) and references therein. As pointed out by Wang and Sun (2007), for missing problems, the inverse probability weighting approach usually depends on high dimensional smoothing for estimating the completely unknown propensity score function, leading to the well known problem of "curse of dimensionality" that may restrict to use of the resulting estimator. One way to avoid such a problem is to use the inverse marginal probability weighted method suggested by Wang et al. (2004).

4.3 Estimation

In model (4.1), consider the case where some values of Y in the sample of size n may be missing, but X and T are fully observed. That is, we obtain the following incomplete

observations

$$(Y_i, \delta_i, X_i, T_i), \qquad i = 1, 2, ..., n$$

from (4.1), where X_i's and T_i's are observed, and,

$$\delta_i = \begin{cases} 0, & Y_i \text{ is missing;} \\ 1, & \text{otherwise.} \end{cases}$$

We assume that Y is missing at random (MAR). The MAR assumption implies that δ and Y are conditionally independent given X and T. That is $P(\delta = 1|Y, X, T) = P(\delta = 1|X, T)$. As discussed above, this is a common assumption for statistical analysis with missing data and is reasonable in many practical situations, see Little and Rubin (1987).

Let us introduce the following notations: $Z = (X, T)$, $\sigma^2(Z) = E(\varepsilon^2|Z)$, $\Delta(z) = P(\delta = 1|Z = z)$ and $\Gamma(t) = P(\delta = 1|T = t)$. Consider the following rank objective function,

$$D_n^C(\beta) = \frac{1}{n} \sum_{i=1}^{n} \varphi\Big(\frac{R(e_i(\beta))}{n+1}\Big) e_i(\beta), \tag{4.2}$$

where $e_i(\beta) = \delta_i \varepsilon_i$ and $R(e_i(\beta))$ is the i^{th} rank of $e_i(\beta)$. Note that in the expression of $D_n^C(\beta)$ in (4.2), β and g are unknown. Also, $E[e_i(\beta)|Z_i] = 0$ by the MAR assumption.

So, before dealing with the estimation of β_0, let us consider first step of the estimation of g based on the complete data, that is, estimating g as a known function of t but unknown with respect to β. As discussed in Wang et al., pre-multiplying (4.1) by the observation indicator, we have

$$\delta_i Y_i = \delta_i X_i^\tau \beta + \delta_i g(T_i) + \delta_i \varepsilon_i,$$

and taking conditional expectations given T, we have

$$E[\delta_i Y_i|T_i = t] = E[\delta_i X_i|T_i = t]\beta + E[\delta_i|T_i = t]g(t), \tag{4.3}$$

from which, it follows that

$$g_1^C(t) = \frac{E[\delta X|T=t]}{E[\delta|T=t]} \quad and \quad g_2^C(t) = \frac{E[\delta Y|T=t]}{E[\delta|T=t]}.$$

and so

$$g(t) = g_2^C(t) - g_1^C(t)\beta. \tag{4.4}$$

Let $K(\cdot)$ be a kernel function and b_n be a bandwidth sequence such that $b_n \to 0 \ \ as \ \ n \to \infty$. Define weights as

$$W_{nj}^C(t) = \frac{K(\frac{t-T_j}{b_n})}{\sum_{j=1}^n \delta_j K(\frac{t-T_j}{b_n})}.$$

Then $\tilde{g}_{1n}^C(t) = \sum_{j=1}^n \delta_j W_{nj}^C(t) X_j$ and $\tilde{g}_{2n}^C(t) = \sum_{j=1}^n \delta_j W_{nj}^C(t) Y_j$ are strongly consistent estimators of $g_1^C(t)$ and $g_2^C(t)$, respectively. Now, define $\widetilde{D}_n^C$ by

$$\widetilde{D}_n^C(\beta) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\nu_{ni}(\beta) \tag{4.5}$$

where $\nu_{in}(\beta) = \delta_i[(Y_i - \tilde{g}_{2n}^C(T_i)) - (X_i - \tilde{g}_{1n}^C(T_i))^\tau \beta]$. Clearly, from the fact that $\tilde{g}_{1n}^C(t) = \sum_{j=1}^n \delta_j W_{nj}^C(t) X_j$ and $\tilde{g}_{2n}^C(t) = \sum_{j=1}^n \delta_j W_{nj}^C(t) Y_j$ are strongly consistent estimators of $g_1^C(t)$ and $g_2^C(t)$ respectively, it can be shown that $\nu_{in}(\beta) \to e_i(\beta)$ in distribution as $n \to \infty$. Define the rank estimator based on complete data as

$$\widetilde{\beta}_\varphi^C = \underset{\beta\in\mathcal{B}}{\text{Argmin}}\, \widetilde{D}_n^C(\beta)\ .$$

Below, we will study the asymptotic properties of rank estimators of β_0, some of which are defined later. The required assumptions are given and discussed in Section 4.4.

Theorem 8. *Under assumptions* $(J_1) - (J_7)$ *except* (J_4),

$$\lim_{n\to\infty} \sup_{\beta\in\mathcal{B}} |\widetilde{D}_n^C(\beta) - D_n^C(\beta)| = 0$$

Proof. In this proof, L is taken to be an arbitrary positive constant not necessarily the same. By the mean value Theorem, there exists $\alpha_n(i)$ such that

$$\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big) - \varphi\Big(\frac{R(e_i(\beta))}{n+1}\Big) = \varphi'(\alpha_n(i))\Big(\frac{R(\nu_{ni}(\beta))}{n+1} - \frac{R(e_i(\beta))}{n+1}\Big)$$

$$\begin{aligned}
\Big|\widetilde{D}_n^C(\beta) - D_n^C(\beta)\Big| &= \Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\nu_{ni}(\beta) - \varphi\Big(\frac{R(e_i(\beta))}{n+1}\Big)e_i(\beta)\Big]\Big| \\
&= \Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\nu_{ni}(\beta) \\
&+ \{\varphi'(\alpha_n(i))\Big(\frac{R(\nu_{ni}(\beta))}{n+1} - \frac{R(e_i(\beta))}{n+1}\Big) - \varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\}e_i(\beta)\Big]\Big|
\end{aligned}$$

Also, by uniform continuity of φ (since φ' is bounded with bound $L > 0$) and for fix n, one can choose $\epsilon(n) = \frac{\xi(n)}{L+1} > 0$, such that for $\xi(n) \to 0$ as $n \to \infty$ and

$$\Big|\frac{R(\nu_{ni}(\beta))}{n+1} - \frac{R(e_i(\beta))}{n+1}\Big| \leq \xi(n) \Longrightarrow \Big|\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big) - \varphi\Big(\frac{R(e_i(\beta))}{n+1}\Big)\Big| < \epsilon(n).$$

From this, we have,

$$\begin{aligned}
\Big|\widetilde{D}_n^C(\beta) - D_n^C(\beta)\Big| &\leq \frac{1}{n}\sum_{i=1}^{n}\Big|\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\Big||\nu_{ni}(\beta) - e_i(\beta)| \\
&+ \frac{1}{n}\sum_{i=}^{n}|\varphi'(\alpha_n(i))|\Big|\frac{R(\nu_{ni}(\beta))}{n+1} - \frac{R(e_i(\beta))}{n+1}\Big||e_i(\beta)|
\end{aligned}$$

From the boundedness of φ, there exists a constant $L > 0$ such that $\Big|\varphi\Big(\frac{R(\nu_{ni}(\beta))}{n+1}\Big)\Big| \leq L$. Note that $\nu_{ni}(\beta) - e_i(\beta) = \delta_i[g_n(T_i) - g(T_i)]$ where $g_n(t) = \tilde{g}_{2n}^C(t) - (\tilde{g}_{1n}^C(t))^\tau\beta$. Also, note that $\sup_{\beta\in\mathcal{B}}|g_n(t) - g(t)| \to 0$ as $n \to 0$. This can be seen from the fact that $g_n(t) - g(t) = (\tilde{g}_{2n}^C(t) - g_2^C(t)) - (\tilde{g}_{1n}^C(t) - g_1^C(t))^\tau\beta$ and then,

$$\sup_{\beta\in\mathcal{B}}|g_n(t) - g(t)| \leq |\tilde{g}_{2n}^C(t) - g_2^C(t))| + \sup_{\beta\in\mathcal{B}}\|\beta\|\|\tilde{g}_{1n}^C(t) - g_1^C(t))\| \to 0 \tag{4.6}$$

since $\tilde{g}_{2n}^C(t) - g_2^C(t) \to 0$ and $\tilde{g}_{1n}^C(t) - g_1^C(t) \to 0$ w.p. 1 as $n \to \infty$. This implies that

$$\sup_{\beta\in\mathcal{B}} \left| \widetilde{D}_n^C(\beta) - D_n^C(\beta) \right| \leq \frac{L}{n} \sum_{i=1}^{n} \sup_{\beta\in\mathcal{B}} |\nu_{ni}(\beta) - e_i(\beta)| + \xi(n) \frac{L}{n} \sum_{i=}^{n} \sup_{\beta\in\mathcal{B}} |e_i(\beta)|$$

Now,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \sup_{\beta\in\mathcal{B}} |\nu_{ni}(\beta) - e_i(\beta)| &\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\beta\in\mathcal{B}} |g_n(T_i) - g(T_i)| \\ &= \int_0^1 \sup_{\beta\in\mathcal{B}} |g_n(t) - g(t)| dm_n(t) \end{aligned}$$

where $m_n(t) = \sum_{i=1}^{n} I(T_i \leq t)$. Applying the Dominated Convergence Theorem, it is easily seen that

$$\int_0^1 \sup_{\beta\in\mathcal{B}} |g_n(t) - g(t)| dm_n(t) \to 0 \quad as \quad n \to \infty.$$

On the other hand, by the strong law of large numbers,

$$\frac{1}{n} \sum_{i=}^{n} \sup_{\beta\in\mathcal{B}} |e_i(\beta)| \to E\Big[\sup_{\beta\in\mathcal{B}} |e| \Big] < \infty.$$

Therefore,

$$\lim_{n\to\infty} \sup_{\beta\in\mathcal{B}} |\widetilde{D}_n^C(\beta) - D_n^C(\beta)| = 0$$

□

Remark 8. *This theorem implies that in the neighborhood of the true parameter β_0, $\widetilde{\beta}_\varphi^C$ and the minimizer of $D_n^C(\beta)$ are equivalent. Under some mild-conditions, $\tilde{\beta}_\varphi^C$ is a strongly consistent estimator of β_0; that is, $\tilde{\beta}_\varphi^C \to \beta_0$ w.p.1. Also, $\tilde{\beta}_\varphi^C$ is a $\sqrt{n}$-consistent estimator of β_0, that is, $\sqrt{n}(\tilde{\beta}_\varphi^C - \beta_0) = O_p(1)$. For reference to these facts, please see Hettmansperger and McKean (1998).*

For $i = 1, \ldots, n$, by imputation ($j = 1$) and inverse probability ($j = 2$), define $\widetilde{Y}_{ijn}$ by

$$\widetilde{Y}_{ijn} = \begin{cases} \delta_i Y_i + \big(1 - \delta_i\big)\big(X_i^\tau \widetilde{\beta}_\varphi^C + \widehat{g}_n^C(T_i)\big), & \text{if } j = 1; \\ \frac{\delta_i}{\widehat{\Gamma}(T_i)} Y_i + \big(1 - \frac{\delta_i}{\widehat{\Gamma}(T_i)}\big)\big(X_i^\tau \widetilde{\beta}_\varphi^C + \widehat{g}_n^C(T_i)\big), & \text{if } j = 2, \end{cases}$$

where, $\widehat{g}_n^C(t) = \tilde{g}_{2n}^C(t) - (\tilde{g}_{1n}^C(t))^\tau \widetilde{\beta}_\varphi^C$, $\widehat{\Gamma}(t) = \sum_{j=1}^{n} \omega_{nj}(t)\delta_j$, with

$$\omega_{nj}(t) = \frac{\Omega(\frac{t-T_j}{h_n})}{\sum_{j=1}^{n} \Omega(\frac{t-T_j}{h_n})} \ .$$

Ω is a kernel function and h_n a bandwidth sequence satisfying $h_n \to 0$ as $n \to \infty$. If we set,

$$Y_{ij} = \begin{cases} \delta_i Y_i + \big(1 - \delta_i\big)\big(X_i^\tau \beta + g(T_i)\big), & \text{if } j = 1; \\ \frac{\delta_i}{\Delta(Z_i)} Y_i + \big(1 - \frac{\delta_i}{\Delta(Z_i)}\big)\big(X_i^\tau \beta + g(T_i)\big), & \text{if } j = 2, \end{cases} \tag{4.7}$$

under the MAR assumption,

$$E(Y_{i1}|Z_i) = E[\delta_i Y_i + \big(1 - \delta_i\big)\big(X_i^\tau \beta + g(T_i)\big)|Z_i] = X_i^\tau \beta + g(T_i).$$

Thus $Y_{i1} = X_i^\tau \beta + g(T_i) + e_i$, with $E(e_i|Z_i) = 0$. Also, for the inverse probability case and under the MAR condition, we have

$$E\Big[\frac{\delta_i}{\Delta(Z_i)} Y_i + \Big(1 - \frac{\delta_i}{\Delta(Z_i)}\Big)\Big(X_i^\tau \beta + g(T_i)\Big)\Big|Z_i\Big] = E\Big[\frac{\delta_i}{\Gamma(T_i)} Y_i + \Big(1 - \frac{\delta_i}{\Gamma(T_i)}\Big)\Big(X_i^\tau \beta + g(T_i)\Big)\Big|Z_i\Big]$$

and

$$E\Big[\frac{\delta_i}{\Gamma(T_i)} Y_i + \Big(1 - \frac{\delta_i}{\Gamma(T_i)}\Big)\Big(X_i^\tau \beta + g(T_i)\Big)\Big|Z_i\Big] = X_i^\tau \beta + g(T_i) \ .$$

Thus $Y_{i2} = X_i^\tau \beta + g(T_i) + \eta_i$, with $E(\eta_i|Z_i) = 0$.

Also, taking the conditional expectation of (4.1) with respect to T, we have

$$E[Y_i|T_i = t] = E[X_i^\tau|T_i = t]\beta + g(t), \tag{4.8}$$

Let $g_1(t) = E[X|T = t]$ and $g_2^j(t) = E[Y|T = t] = E[Y_{ij}|T = t]$, $j = 1, 2$. This suggests that

$$g(t) = g_2(t) - g_1(t)^\tau\beta. \tag{4.9}$$

Let

$$W_{nj}(t) = \frac{M(\frac{t-T_j}{k_n})}{\sum_{j=1}^n M(\frac{t-T_j}{k_n})}.$$

where M is a kernel function and k_n is a bandwidth sequence ($k_n \to 0$ as $n \to \infty$). Define $\hat{g}_{1n}(t) = \sum_{k=1}^n W_{nk}(t)X_k$ and $\hat{g}_{2n}^j(t) = \sum_{k=1}^n W_{nk}(t)\widetilde{Y}_{kjn}$. Under some mild conditions, it can be shown that $\hat{g}_{1n}(t) \to g_1(t)$ and $\hat{g}_{2n}^j(t) \to g_2^j(t)$ with probability 1.

Theorem 9. *Under assumptions* $(J_5) - (J_9)$ *and Remark 8*

$$|g_n(t) - \hat{g}_n^C(t)| \to 0 \quad as \quad n \to \infty \quad w.p\ 1.$$

Proof.

$$\begin{aligned} |g_n(t) - \hat{g}_n^C(t)| &= |\tilde{g}_n(t) - g(t) + g(t) - \hat{g}_n^C(t)| \\ &\leq |g_n(t) - g(t)| + |g(t) - \hat{g}_n^C(t)| \end{aligned}$$

Clearly, from (4.6), $|g_n(t) - g(t)| \to 0$ w.p.1 as $n \to \infty$. Then, to complete the proof, we just need to show that $|g(t) - \hat{g}_n^C(t)| \to 0$ w.p.1 as $n \to \infty$. Indeed,

$$\begin{aligned} g(t) - \hat{g}_n^C(t) &= g_2^C(t) - (g_1^C(t))^\tau\beta - \left(\tilde{g}_{2n}^C(t) - (\tilde{g}_{1n}^C(t))^\tau\beta_0\right) \\ &= (g_2^C(t) - \tilde{g}_{2n}^C(t)) + (\tilde{g}_{1n}^C(t) - \tilde{g}_1^C(t))^\tau\beta_0 + (\tilde{g}_1^C(t))^\tau(\tilde{\beta}_\varphi^C - \beta_0) \end{aligned}$$

Then, $|g(t) - \hat{g}_n^C(t)| \leq |g_2^C(t) - \tilde{g}_{2n}^C(t)| + \|\tilde{g}_{1n}^C(t) - \tilde{g}_1^C(t)\|\|\beta_0\| + \|\tilde{g}_{1n}^C(t)\|\|\tilde{\beta}_\varphi^C - \beta_0\|$. Clearly, $|g_2^C(t) - \tilde{g}_{2n}^C(t)| \to 0$ and $|g_1^C(t) - \tilde{g}_{1n}^C(t)| \to 0$ w.p.1 by consistency. $\tilde{g}_{1n}^C(t)$ is bounded since it converges. Then, $\|\tilde{g}_{1n}^C(t)\|\|\tilde{\beta}_\varphi^C - \beta_0\| \to 0$ w.p.1 since $\tilde{\beta}_\varphi^C - \beta_0 \to 0$ w.p.1. as $n \to \infty$. Therefore,

$$|g_n(t) - \hat{g}_n^C(t)| \to 0 \;\; as \;\; n \to \infty \;\; w.p\ 1.$$

□

Set

$$\mathbf{v}_n^{ij}(\beta) = \widetilde{Y}_{ijn} - \hat{g}_{2n}^j(T_i) - \big(X_i - \hat{g}_{1n}(T_i)\big)^\tau \beta \ .$$

Now, based on simple imputation, the rank estimator $\widehat{\beta}_\varphi^I$ of β_0 is defined by $\widehat{\beta}_\varphi^I = \underset{\beta\in\mathcal{B}}{\text{Argmin}}\, D_n^I(\beta)$ with $D_n^I(\beta)$ defined by

$$D_n^I(\beta) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R^I(\mathbf{v}_n^{i1}(\beta))}{n+1}\Big)\mathbf{v}_n^{i1}(\beta) \ ,$$

where $R^I(\mathbf{v}_n^{i1}(\beta))$ is the i^{th} rank of $\mathbf{v}_n^{i1}(\beta)$. Now letting $S_n^I(\beta) = -\nabla D_n^I(\beta)$, we have

$$S_n^I(\beta) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R^I(\mathbf{v}_n^{i1}(\beta))}{n+1}\Big)\big(X_i - \tilde{g}_{1n}(T_i)\big). \tag{4.10}$$

On the other hand, based on the inverse probability, the rank estimator $\widehat{\beta}_\varphi^{IP}$ of β_0, is defined by $\widehat{\beta}_\varphi^{IP} = \underset{\beta\in\mathcal{B}}{\text{Argmin}}\, D_n^{IP}(\beta)$ with

$$D_n^{IP}(\beta) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R^{IP}(\mathbf{v}_n^{i2}(\beta))}{n+1}\Big)\mathbf{v}_n^{i2}(\beta) \ ,$$

where $R^{IP}(\mathbf{v}_{in}^2(\beta))$ is the i^{th} rank of $\mathbf{v}_n^{i2}(\beta)$. Letting $S_n^{IP}(\beta) = -\nabla D_n^{IP}(\beta)$, we have

$$S_n^{IP}(\beta) = \frac{1}{n}\sum_{i=1}^n \varphi\Big(\frac{R^{IP}(\mathbf{v}_n^{i2})}{n+1}\Big)\big(X_i - \tilde{g}_{1n}(T_i)\big) \tag{4.11}$$

Note that $\hat{g}_{1n}(T_i)$ is a p-vector and the estimation of g_2^j, involves all the Y_i's, making $\{\mathbf{v}_n^{ij}(\beta_0)\}_{i=1}^n$ to be a set of dependent random variables. Also if we set G_{in}^1 and G_{in}^2 to denote distribution functions of $\mathbf{v}_n^{i1}(\beta_0)$ and $\mathbf{v}_n^{i2}(\beta_0)$ respectively, it can be easily shown that

$$G_i^1(s) = \lim_{n\to\infty} G_{in}^1(s) = \lim_{n\to\infty} P(\mathbf{v}_n^{i1}(\beta_0) \leq s)$$

and

$$G_i^2(s) = \lim_{n\to\infty} G_{in}^2(s) = \lim_{n\to\infty} P(\mathbf{v}_n^{i2}(\beta_0) \leq s),$$

where G_i^1 and G_i^2 are conditional distribution functions of e_i and η_i given Z_i, respectively.

4.4 Assumptions

The following assumptions were used above to motivate the theory leading to the definition of the rank estimators $\widehat{\beta}_\varphi^I$ and $\widehat{\beta}_\varphi^{IP}$. They will be useful in studying the asymptotic properties of the rank estimators $\widehat{\beta}_\varphi^I$ and $\widehat{\beta}_\varphi^{IP}$.

(J_1) $P(X^\tau\beta = X^\tau\beta_0) < 1$ for any $\beta \neq \beta_0$,

(J_2) φ is a bounded, twice continuously differentiable score function with bounded derivatives, defined on $(0,1)$, and, satisfying:

$$\int_0^1 \varphi(u)du = 0 \quad and \quad \int_0^1 \varphi^2(u)du = 1$$

(J_3) The cumulative distribution H of Y given Z is symmetric about 0 and has a corresponding density h that is absolutely continuous with finite Fisher information, i.e., $I(h) = \int_{-\infty}^{\infty} \left[\frac{h'(\varepsilon)}{h(\varepsilon)}\right]^2 h(\varepsilon)d\varepsilon < \infty$. Define $\gamma_\varphi = \int_0^1 \varphi(u)\varphi_H(u)du$, where $\varphi_H(u) = \frac{h'(H^{-1}(u))}{h(H^{-1}(u))}$.

(J_4) Define $X^* = X - \tilde{g}_{1n}(T)$. Assume that $\frac{1}{n}X^\tau X \to \Sigma$, $\frac{1}{n}\tilde{\mathbf{G}}_n^\tau \tilde{\mathbf{G}}_n \to \mathbf{G}$ and $\frac{1}{n}\big(\tilde{\mathbf{G}}_n^\tau X + X^\tau \tilde{\mathbf{G}}_n\big) \to \mathbf{B}$, for some positive definite matrices Σ, $\mathbf{G}$ and $\mathbf{B}$. Also, assume that

$$P(\|X^*\| \geq c_n) = o_P(d_n) \quad for \quad d_n \to 0 \quad as \quad c_n \to \infty.$$

(J_5) $\inf_t \Gamma(t) > 0$, $\inf_z \Delta(z) > 0$ and the density of T, say $m(t)$, exits and satisfies

$$0 < \inf_t m(t) \leq \sup_t m(t) < \infty.$$

Also $\Gamma(\cdot)$ has bounded derivatives.

(J_6) $g_1^C(\cdot)$, $g_2^C(\cdot)$, $g_1(\cdot)$, and $g_2^j(\cdot)$ have continuous derivatives at t and bounded derivatives up to order 2.

(J_7) The kernels $K(\cdot)$, $\Omega(\cdot)$ and $M(\cdot)$ are regular kernels of order $r(> 2)$.

(J_8) The bandwidth b_n, h_n and k_n satisfy the following conditions:

i.) $nk_nb_n \to \infty$, $nb_n^{4r} \to 0$, $nk_n^{4r} \to 0$ and $b_n^2/k_n \to 0$

ii.) $nh_n \to \infty$ and $nh_n^{4r} \to 0$

iii.) $C(\log n/n)^\gamma < h_n,\ b_n,\ k_n < \zeta_n$, for any $C > 0$, $\gamma = 1 - 2/p$, $p > 2$ and ζ_n not necessarily the same for the three bandwidths such that $C(\log n/n)^\gamma < \zeta_n < 1$ satisfying $\zeta_n \to 0$ as $n \to \infty$.

(J_9) $\sup_t E[\|X\|^p | T = t] < \infty$ and $\sup_t E[|Y|^p | T = t] < \infty$ for $p > 2$.

Remark 9. *Assumption* (J_1) *stands for the identifiability condition and together with* (J_2)*, ensures the consistency of the resulting estimator.* (J_5)*,* (J_6)*,* (J_7)*,* (J_8) *and* (J_9) *are standard assumptions for nonparametric regression problems. Specifically* (J_5) *ensures the non missingnesss with probability* 1 *anywhere in the domain of* Z*. Assumption iii.) in* (J_8) *and*

(J_9) ensure the uniform consistency of the Nadaraya Watson estimator used to the true unknown function g, for more discussion see Einmahl and Mason (2005). As pointed out by Xue (2009), the assumption $P(\|X^*\| \geq c_n) = o_p(d_n)$ for $d_n \to 0$ as $c_n \to \infty$ in (J_4) is commonly used for avoiding the boundary problem. This assumption is also used by Zhu and Fang (1996) and Wang and Rao (2002a). For a class of distributions that satisfy this assumption, see discussion in Xue (2009). (J_1), (J_2), (J_3) and (J_4) together with others assumptions listed above, ensure the asymptotic normality of the resulting estimator. For practical issue, the optimal bandwidth can be chosen to lie in the interval $[a.n^{-1/5}, b.n^{-1/5}]$ for $0 < a < b < \infty$. See Einmahl and Mason (2005) for more discussion. Most of regular kernels in (J_7) satisfy assumption $(C6)$ of Xue (2009) or Wang and Sun (2007).

4.5 Asymptotic Normality

Put $X = (X_1, \cdots, X_n)^\tau$ and $\tilde{\mathbf{G}}_n = (\tilde{g}_{1n}(T_1), \cdots, \tilde{g}_{1n}(T_n))^\tau$ two matrices given by

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ X_{21} & \cdots & X_{2n} \\ & & \\ X_{p1} & \cdots & X_{pn} \end{pmatrix} \quad \tilde{\mathbf{G}}_n = \begin{pmatrix} \tilde{g}_{11n}(T_1) & \cdots & \tilde{g}_{1pn}(T_1) \\ \tilde{g}_{11n}(T_2) & \cdots & \tilde{g}_{1pn}(T_2) \\ & & \\ \tilde{g}_{11n}(T_n) & \cdots & \tilde{g}_{1pn}(T_n) \end{pmatrix}.$$

For simplicity, set $D_n^l = D_n^I$, $S_n^l = S_n^I$ for $l = I$ and $D_n^l = D_n^{IP}$, $S_n^l = S_n^{IP}$ for $l = IP$. Now, as discussed in Hettmansperger and McKean (1998), if we set

$$M_n^l(\beta) = (2\gamma_\varphi)^{-1} X^{*\tau} X^* (\beta - \beta_0) - (\beta - \beta_0)^\tau S_n^l(\beta_0) + D_n^l(\beta_0),$$

we obtain the following result called asymptotic quadraticity which was proved by Jaeckel (1972).

Theorem 10. *Under assumptions* $(J_1) - (J_6)$, $\forall \epsilon > 0$ *and* $C > 0$,

$$\lim_{n\to\infty} P_{\beta_0}\Big[\max_{\|\beta-\beta_0\|\leq \frac{C}{\sqrt{n}}} |D_n^l(\beta) - M_n^l(\beta)| \geq \epsilon\Big] = 0$$

This result provides a quadratic approximation of D_n^l by M_n^l and leads to the asymptotic linearity derived by Jureckova (1971) and is displayed as follows.

Theorem 11. *Under* $(J_1) - (J_9)$, *for any* $C > 0$ *and* $\epsilon > 0$,

$$\lim_{n\to\infty} P_{\beta_0}\Big[\sup_{\|\beta-\beta_0\|\leq n^{-1/2}C} \Big\|n^{-1/2}[S_n^l(\beta) - S_n^l(\beta_0)] + n^{-1/2}\gamma_\varphi^{-1}X^{*\tau}X^*(\beta-\beta_0)\Big\| \geq \epsilon\Big] = 0$$

Proofs of Theorem 10 and Theorem 11 can be constructed in a straightforward manner along the lines discussed in Hettmansperger and McKean (1998) for the linear model setting. Therefore, for the sake of brevity, they will not be included here.

Theorem 12. *Under assumptions* $(J_1) - (J_9)$, $\sqrt{n}S_n^l(\beta_0) \approx N_p(0, \Sigma_{jn}) \xrightarrow{\mathcal{D}} N_p(0, \mathbf{V}_j)$ *where* $j = 1$ *for* $l = I$ *and* $j = 2$ *for* $l = IP$. $\mathbf{V}_j = \lim_{n\to\infty} \Sigma_{jn}$.

Before giving the proof of Theorem 12, consider the following notation from Brunner and Denker (1994). Set

$$\boldsymbol{\lambda}_{in} = X_i - \hat{g}_{1n}(T_i), \quad J_{jn}(s) = \frac{1}{n}\sum_{i=1}^n G_{in}^j(s), \quad \hat{J}_{jn}(s) = \frac{1}{n}\sum_{i=1}^n I(\mathbf{v}_n^{i1}(\beta_0) \leq s),$$

$$F_{jn}(s) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_{in} G_{in}^j(s), \quad \hat{F}_{jn}(s) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_{in} I(\mathbf{v}_n^{ij}(\beta_0) \leq s)$$

and $T_n^l(\beta_0) = \dfrac{S_n^l(\beta_0)}{\alpha(n)} - E\Big[\dfrac{S_n^l(\beta_0)}{\alpha(n)}\Big]$.

Lemma 5 (Brunner and Denker (1994))**.** *Suppose that* $M_0 n^\gamma \leq m_1(n) \leq M_1 n^\gamma$ *for some constants* $0 < M_0 \leq M_1 < \infty$ *and* $\gamma > 0$, *and that* $\zeta_{jn} \geq Cn^a$ *for some constant* a, $C \in \mathbb{R}$,

where

$$U_{jn} = \int \varphi(J_{jn}(s))(\hat{F}_{jn} - F_{jn})(ds) + \int \varphi'(J_{jn}(s))(\hat{J}_{jn}(s) - J_{jn}(s))F_{jn}(ds).$$

and ζ_{jn} is the minimum eigenvalue of $Var(U_{jn})$. Then $n\mathbf{W}_{jn}^{-1}T_n^l(\beta_0)$ is asymptotically standard multivariate normal, provided φ is twice continuously differentiable, with bounded second derivative and $\gamma < (a+1)/2$.

Proof of Theorem 12. Assume that $\max_{1\leq i\leq n} \|\boldsymbol{\lambda}_{in}\| = \alpha(n) < \infty$. From the setting defined above,

$$S_n^l(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} \varphi\Big(\frac{R^l(\mathbf{v}_n^{ij}(\beta_0))}{n+1}\Big)\big(X_i - \hat{g}_{1n}(T_i)\big) = \int \varphi\Big(\frac{n}{n+1}\hat{J}_{jn}\Big)dF_{jn}$$

Now define

$$\boldsymbol{\Lambda}_n = n^2 E[(T_n^l(\beta_0))(T_n^l(\beta_0))^\tau], \quad B_{jn} = -\int (\hat{F}_{jn} - F_{jn})d\varphi(J_{jn}) + \int (\hat{J}_{jn} - J_{jn})\frac{dF_{jn}}{dJ_{jn}}d\varphi(J_{jn})$$

and $\mathbf{W}_{jn} = n^2 Var(B_{jn})$. From the fact that $\beta_0 = \underset{\beta\in\mathcal{B}}{\text{Argmin}}\, E(D_n^l(\beta))$, it can be seen that $E[S_n^l(\beta_0)] = 0$. This implies that $E[T_n^l(\beta_0)] = 0$ and $T_n^l(\beta_0) = \dfrac{S_n^l(\beta_0)}{\alpha(n)}$. Now under assumption (J_2) and by Brunner and Denker (1994), for Theorem 12 to hold, it suffices to check if conditions of Lemma 5 are satisfied. To that end, the fact that $Var(\varepsilon_i|Z_i) > 0$, there exists $\epsilon > 0$ such that the minimum eigenvalue of $Var(B_{jn})$, say μ_{jn} satisfies $\mu_{jn} > \epsilon n^b$, for $0 < b < 1/2$. This is obtained under the assumption that $\varsigma_{jn}/n \to \infty$ putting $\mu_{jn} \to \infty$ as $n \to \infty$, see Brunner and Denker (1994) for more discussion. Again by Brunner and Denker (1994), $U_{jn} = nB_{jn}$ and then, $Var(U_{jn}) = n^2 Var(B_{jn})$. Thus, $\xi_{jn} = n^2\mu_{jn} \geq \epsilon n^{2+b}$. Hence, putting $a = 2 + b$, $\gamma = 1$, $M_0 = M_1$ and $C = \epsilon$, conditions of Lemma 5 are satisfied. This, shows that $n\mathbf{W}_{jn}^{-1}T_n^l(\beta_0)$ is asymptotically multivariate standard normal. Therefore $\sqrt{n}S_n^l(\beta_0)$ is asymptotically multivariate normal with mean 0 and covariance matrix $\Sigma_{jn} = \dfrac{\alpha(n)}{\sqrt{n}}\mathbf{W}_{jn}$. □

Remark 10. *For $l = I$ or $l = IP$, note that by definition of $\widehat{\beta}_{\varphi}^{l}$, $S_n^l(\widehat{\beta}_{\varphi}^{l}) = 0$. Also, from the fact that M_n^l is a quadratic function of β, it is uniquely minimized by $\tilde{\beta}_{\varphi}^{l} = \gamma_{\varphi}\big(X^{*\tau}X^{*}\big)^{-1}S_n^l(\beta_0)$. Assumption (J_4) ensures that $\frac{1}{n}\big(X^{*\tau}X^{*}\big) \to \Sigma^{*}$ as $n \to \infty$ for some positive definite matrix Σ^{*}. Then, under the assumptions of Theorem 11, $\sqrt{n}(\widehat{\beta}_{\varphi}^{l} - \tilde{\beta}_{\varphi}^{l}) = o_p(1)$. The proof of this claim can also be constructed along the lines given in Hettmansperger and McKean (1998). Moreover, conditioning on Z and using the SLLN, $S_n^I(\beta_0) \to 0$ $w.p.1$.*

The following theorem gives a practical way of estimating the covariance matrix $\Sigma_{jn} = \frac{\alpha(n)}{\sqrt{n}}\mathbf{W}_{jn}$.

Theorem 13. *Suppose that assumptions $(J_1) - (J_9)$ hold and let $j = 1$ for $l = I$ and $j = 2$ for $l = IP$. Define*

$$\begin{aligned}
\widehat{Z}_j^l &= \sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\varphi\Big(\frac{R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l}))}{n+1}\Big) + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\varphi'\Big(\frac{R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l}))}{n+1}\Big)R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l})) \\
&= nS_n^l(\widehat{\beta}_{\varphi}^{l}) + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\varphi'\Big(\frac{R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l}))}{n+1}\Big)R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\varphi'\Big(\frac{R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l}))}{n+1}\Big)R(\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l})),
\end{aligned}$$

since $S_n^l(\widehat{\beta}_{\varphi}^{l}) = 0$. Then, $\widehat{\mathbf{W}}_{jn} = [\widehat{Z}_j^l - E(Z_j^l)][\widehat{Z}_j^l - E(Z_j^l)]^{\tau} \to \mathbf{W}_{jn}$ (positive definite) in the L^2-norm as $n \to \infty$, where

$$\begin{aligned}
Z_j^l &= n\int \varphi(J_{jn}(t))\hat{F}_{jn}(dt) + \int \varphi'(J_{jn}(t))\hat{J}_{jn}(t)F_{jn}(dt) \\
&= nS_n^l(\beta_0) + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\varphi'\Big(\frac{R(\mathbf{v}_n^{ij}(\beta_0))}{n+1}\Big)R(\mathbf{v}_n^{ij}(\beta_0))
\end{aligned}$$

Proof. By Theorem 4.1 of Brunner and Denker (1994), to prove this theorem it is sufficient to prove that $\lim_{n\to\infty} \frac{n}{\varsigma_{jn}} = 0$, where ς_{jn} represents the minimum eigenvalue of $\mathbf{W}_{jn}$. This is obtained from the fact that $\mu_{jn} \geq \epsilon n^b$ for $0 < b < 1/2$. □

Note that $E[S_n^l(\beta_0)] = 0$. Then, conditioning on Z, we have,

$$\begin{aligned} E[Z_j^l] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\lambda}_{in} E\Big[\varphi'\Big(\frac{R(\mathbf{v}_n^{ij}(\beta_0))}{n+1}\Big) R(\mathbf{v}_n^{ij}(\beta_0))\Big] &= \sum_{i=1}^{n} \boldsymbol{\lambda}_{in} \int \varphi'(G_{in}^j(u)) dG_{in}^j(u) \\ &\approx \sum_{i=1}^{n} \boldsymbol{\lambda}_{in} \int \varphi'(G_i^j(u)) dG_i^j(u) \\ &\approx \Big(\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\Big) \int \varphi'(G^j(u)) dG^j(u) \end{aligned}$$

where

$$G^j(u) = \begin{cases} \Delta(z)H(u), & \text{if } j = 1; \\ \frac{\Delta(z)}{\Gamma(t)}H(u), & \text{if } j = 2. \end{cases}$$

For either $j = 1$ or $j = 2$, considering the change of variables, say, $s = \Delta(z)H(u)$ or $s = \dfrac{\Delta(z)}{\Gamma(t)}H(u)$, we get

$$\int \varphi'(G^j(u)) dG^j(u) = \int \varphi'(s) ds$$

Hence,

$$E[Z_j^l] \approx \Big(\sum_{i=1}^{n} \boldsymbol{\lambda}_{in}\Big) \int \varphi'(s) ds$$

Theorem 14. *Under assumptions* $(J_1) - (J_6)$*, we have*

$$\sqrt{n}(\widehat{\beta}_\varphi^l - \beta_0) \xrightarrow{\mathcal{D}} N(0, \gamma_\varphi^2 \Sigma_j^{-1}).$$

where

$$\Sigma_j = \Sigma^{*-1} \mathbf{V}_j \Sigma^{*-1} \, .$$

The proof of this Theorem is obtained by combining results of Theorem 11, Lemma 12 and the discussion in Remark 10.

4.6 Estimation of the function g

As it can seen from the discussion in section 2, function g is not fully estimated. In estimating β_0, the first step of the estimation of g was used by setting it as a known function of t but unknown as a function of β throughout the estimates of g_1 and g_2^j. As soon as the estimated values $\widehat{\beta}_\varphi$ of β_0 are obtained, one can now obtain the estimated function $\hat{g}_n$ of g accordingly to the estimated value of β_0 by putting:

$$\hat{g}_n(t) = \hat{g}_{2n}^j(t) - \hat{g}_{1n}(t)\widehat{\beta}_\varphi^l$$

where $j = 1$ for $l = I$ and $j = 2$ for $l = IP$.

Theorem 15. *Under assumptions* $(J_1) - (J_9)$*, we have* $\hat{g}_n(t) - g(t) = o(1)$ *with probability 1.*

Proof. By definition of $\hat{g}_n(t)$, we have

$$\hat{g}_n(t)-g(t) = \hat{g}_{2n}^j(t)-g_2(t)-\big(\hat{g}_{1n}(t)-g_1(t)\big)^\tau(\widehat{\beta}_\varphi-\beta_0)-(g_1(t))^\tau(\widehat{\beta}_\varphi-\beta_0)-\big(\hat{g}_{1n}(t)-g_1(t)\big)^\tau\beta_0$$

This implies that,

$$|\hat{g}_n(t)-g(t)| \leq |\hat{g}_{2n}^j(t)-g_2(t)|+\|\hat{g}_{1n}(t)-g_1(t)\|\|\widehat{\beta}_\varphi-\beta_0\|+\|(g_1(t))\|\|\widehat{\beta}_\varphi-\beta_0\|+\|\hat{g}_{1n}(t)-g_1(t)\|\|\beta_0\|$$

We continue the proof with $j = 1$ and similar argument can be used to derive that of $j = 2$. By definition of $\hat{g}^1_{2n}(t)$,

$$\begin{aligned}
\hat{g}^1_{2n}(t) - g_2(t) &= \sum_{i=1}^{n} W_{ni}(t)\widetilde{Y}_{i1n} - g_2(t) \\
&= \sum_{i=1}^{n} W_{ni}(t)[\delta_i Y_i + \big(1-\delta_i\big)\big(X_i^\tau \widetilde{\beta}^C_\varphi + \widehat{g}^C_n(T_i)\big) - g_2(t)] \\
&= \sum_{i=1}^{n} W_{ni}(t)(Y_{i1} - g_2(t)) + \sum_{i=1}^{n} W_{ni}(t)(1-\delta_i)X_i^\tau(\tilde{\beta}^C_\varphi - \beta_0) \\
&+ \sum_{i=1}^{n} W_{ni}(t)(1-\delta_i)(\hat{g}^C_n - g(t)).
\end{aligned}$$

Note that $E[Y_{i1}|T_i = t] = g_2(t)$ and $E[\|(1-\delta_i)X_i\||T_i] < \infty$. Then, under mild-conditions, we have, $|\sum_{i=1}^{n} W_{ni}(t)Y_{i1} - g_2(t)| \to 0 \quad w.p.1$ and by Theorem 9,

$$|\hat{g}^C_n - g(t)| \to 0 \quad w.p.1.$$

Also, conditioning on T and using the SLLN, we have

$$\sum_{i=1}^{n} W_{ni}(t)(1-\delta_i)X_i \to E[\|(1-\delta)X\||T] \quad w.p.1.$$

Hence, $\sum_{i=1}^{n} W_{ni}(t)(1-\delta_i)X_i = O(1)$. From the fact that $\tilde{\beta}^C_\varphi - \beta_0 = o(1)$, we have

$$\hat{g}^1_{2n}(t) - g_2(t) \to 0 \quad w.p.1.$$

Using the same argument, it can be shown that $\hat{g}_{1n}(t) - g_1(t) \to 0 \quad w.p.1$. Combining these facts with Remark 10 completes the proof. □

4.7 Bandwidth Selection

An important issue when dealing with kernel estimation, is the selection of an appropriate bandwidth sequence. In the nonparametric regression literature, this problem has been studied extensively. The common approach used in selecting the bandwidth, is the delete-one cross-validation rule. This approach consists in minimizing the "leave-one-out" version of objective function evaluated at the estimator of β_0 as a function the bandwidth h. That is, b_n, h_n and k_n may be obtained by minimizing

$$\sum_{i=1}^{n} \varphi\Big(\frac{R(\hat{u}_{-i}(h))}{n+1}\Big)\hat{u}_{-i}(h),$$

where

$$\hat{u}_{-i}(h) = \begin{cases} \nu_{-i,n}(\tilde{\beta}_{\varphi}^{C}, h), & \text{for } b_n; \\ \mathbf{v}_n^{-i,j}(\widehat{\beta}_{\varphi}^{l}, h), & \text{for } k_n \ . \\ \delta_i - \widehat{\Gamma}_{-i}(T_i, h), & \text{for for } h_n. \end{cases}$$

with $\nu_{-i,n}(\tilde{\beta}_{\varphi}^{C}, h)$, $\mathbf{v}_n^{-i,j}(\widehat{\beta}_{\varphi}^{l}, h)$ and $\delta_i - \widehat{\Gamma}_{-i}(T_i, h)$ being the "leave-one-out" versions of $\nu_{i,n}(\tilde{\beta}_{\varphi}^{C})$, $\mathbf{v}_n^{ij}(\widehat{\beta}_{\varphi}^{l})$, $\delta_i - \widehat{\Gamma}(T_i)$, respectively, in which $b_n,\ k_n,\ h_n$ are replaced by h.

4.8 Simulation

To fully understand the optimality properties of the proposed approach, we will consider the finite sample behavior of the estimator. To do so, a simulation study is conducted. The model used in the simulation is given by

$$Y = \beta_0 x + g(T) + \varepsilon$$

with x generated from a normal distribution with mean 1 and variance 1. The random errors ε are generated from the contaminated normal distribution $CN(\gamma, \sigma) = (1-\gamma)N(0,1) +$

$\gamma N(0, \sigma^2)$ with different degrees of contamination and the t-distribution with various degrees of freedom. The kernels $K(\cdot)$ and $M(\cdot)$ were taken to be the Epanechnikov Kernel

$$K(t) = M(t) = 0.75(1 - t^2)I(|t| \leq 1).$$

Based on assumption (J_8), all the three bandwidths b_n, k_n and h_n were taken to be proportional to $n^{-1/5}$. Two simulation scenarios were considered. In Scenario 1, the true g was taken to be 0 and $T = \dfrac{i}{n^2}$, for $i = 1, \cdots, n$ and for $z = (x, t)$, three cases were considered:

- Case 1: $\Delta(z) = 0.8 + 0.2|x - 1|$ if $|x - 1| \leq 1$, and 0.95 elsewhere.
- Case 2: $\Delta(z) = 0.9 - 0.2|x - 1|$ if $|x - 1| \leq 4.5$, and 0.1 elsewhere.
- Case 3: $\Delta(z)$ is taken to be a sequence of proportions of missingness starting from 0% to 50% with steps of 10%.

In Scenario 2, the true g was taken to be $g(t) = (\sin(2\pi t^2))^{1/3}$ and T is generated from the uniform distribution $U[0, 1/4]$. One more case is added to the three previous.

- Case 4: $\Delta(z) = 0.9 - 0.2(|x-1| + |t-0.5|)$ if $|x-1| + |t-0.5| \leq 1.5$, and 0.8 elsewhere.

In all the cases, δ was generated from the Bernoulli$(\Delta(z))$ distribution. Under both scenarios β is estimated using simple imputation **(SI)** and based on the inverse probability procedure where $\Omega(t)$ is chosen in two ways. First, as defined above, we take it to be kernel **(Ker)**, that is $\Omega(t) = M(t)$, and secondly by taking it to be the logistic function **(Log)** given by $\Omega(t) = \dfrac{1}{1 + e^{-t}}$ as in Müller (2009). From 5000 simulations, the MSEs of the resulting rank (R) estimator of β_0 are reported and are compared to those of the least squares (LS) estimator of β_0. Figures 4.1 – 4.17 contain the results of the simulation experiments.

4.9 Results and Discussion

The results of the simulation experiment are described below:

Figure 4.1: Scenario 1, Case 1: MSE vs Proportion of Contamination and t-df

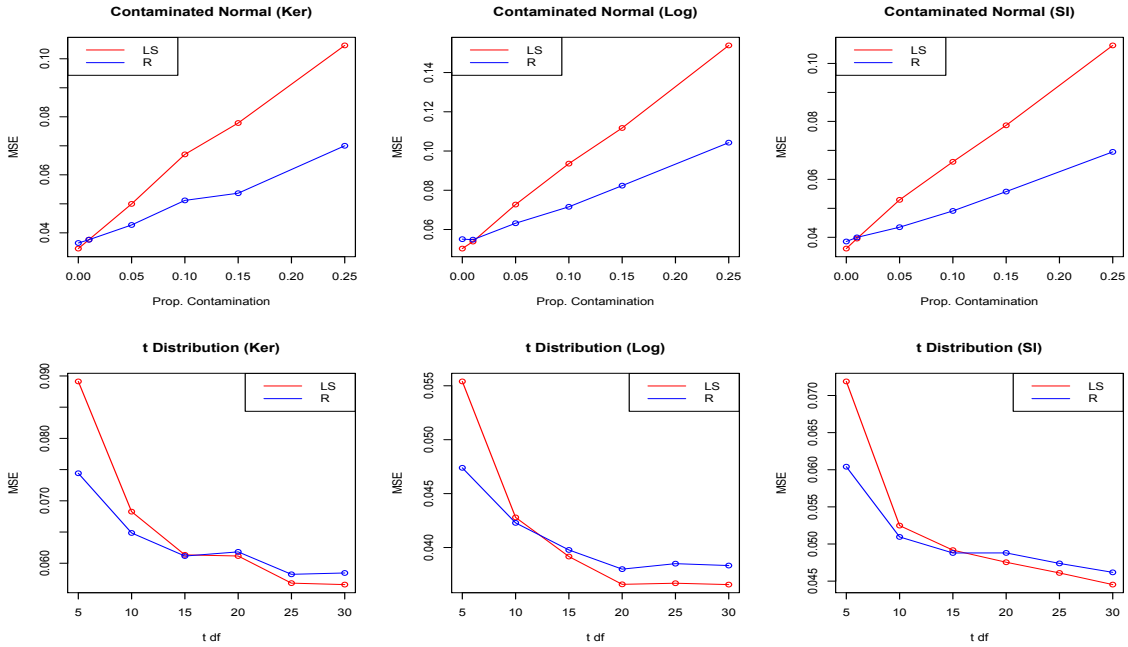


Figure 4.2: Scenario 1, Case 2: MSE vs Proportion of Contamination and t-df

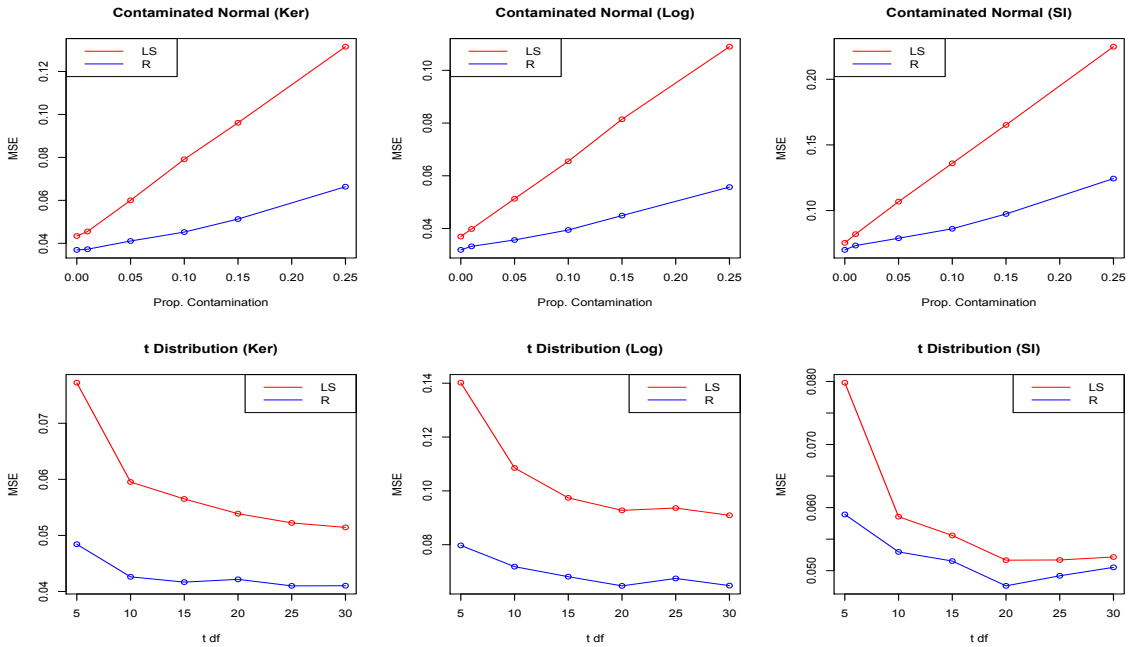


Figure 4.3: Scenario 1, Case 3: MSE vs Proportion of Missing Data for SI under contaminated normal error distribution

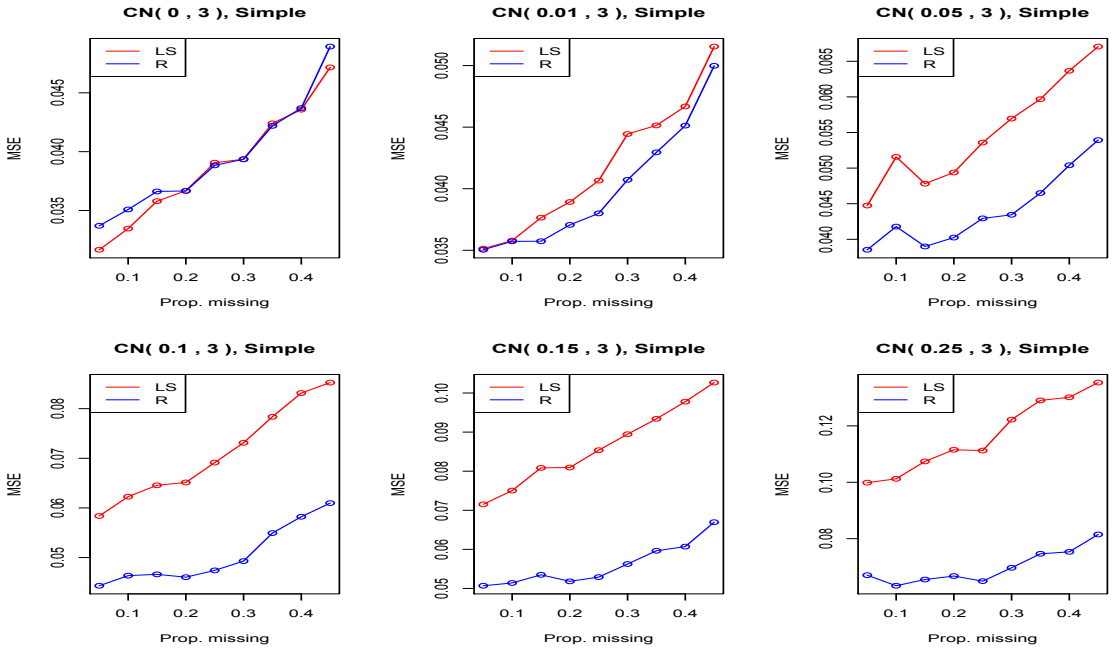


Figure 4.4: Scenario 1, Case 3: MSE vs Proportion of Missing Data for SI under t error distribution

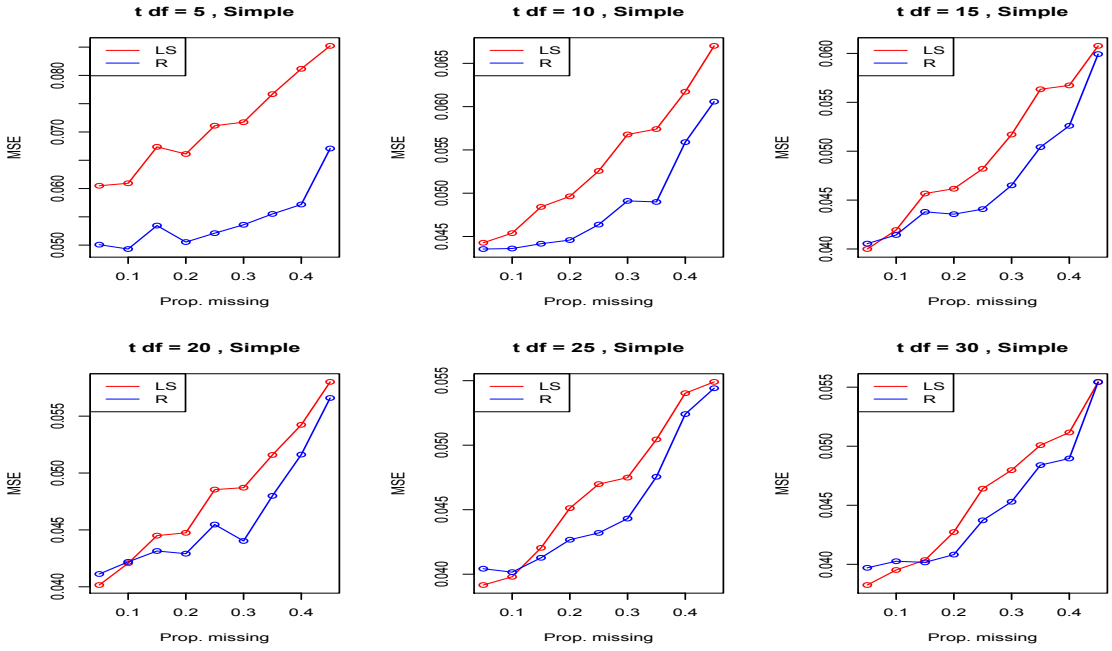


Figure 4.5: Scenario 1, Case 3: MSE vs Proportion of Missing Data for inverse probability (Ker) under contaminated normal error distribution

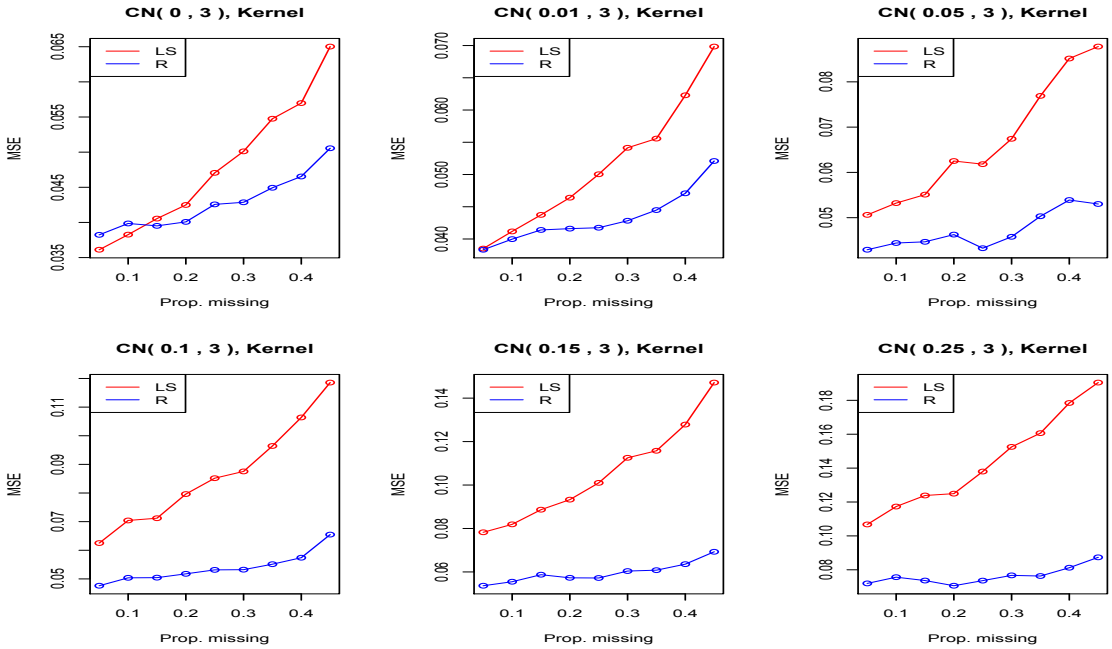


Figure 4.6: Scenario 1, Case 3: MSE vs Proportion of Missing Data for inverse probability (Ker) under t error distribution

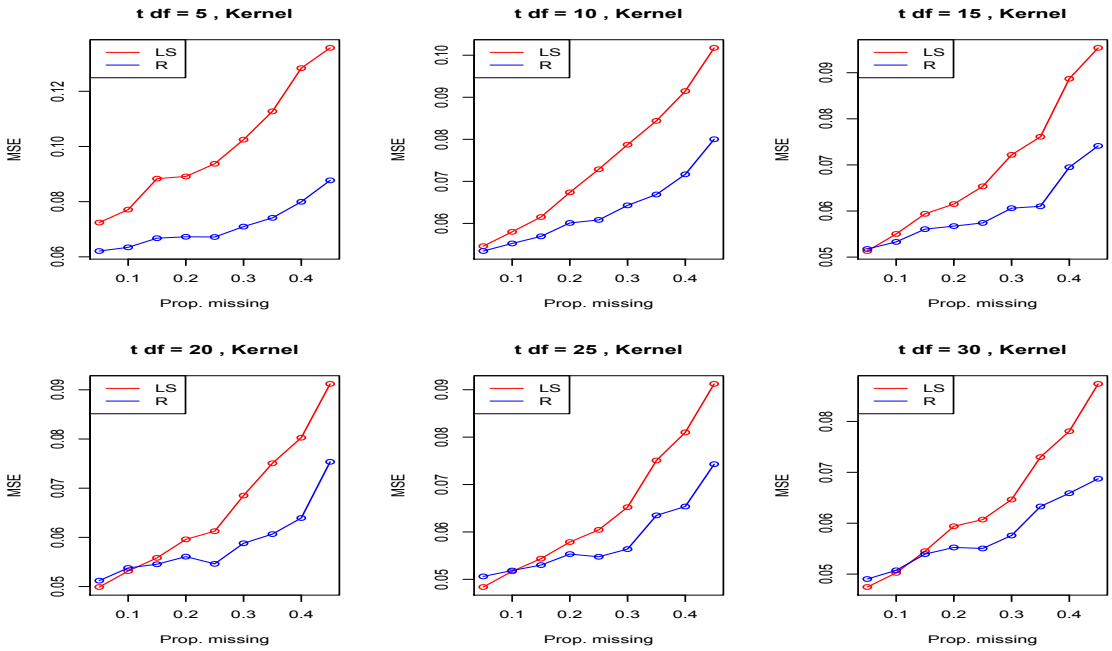


Figure 4.7: Scenario 1, Case 3: MSE vs Proportion of Missing Data for inverse probability (Log) under contaminated normal error distribution

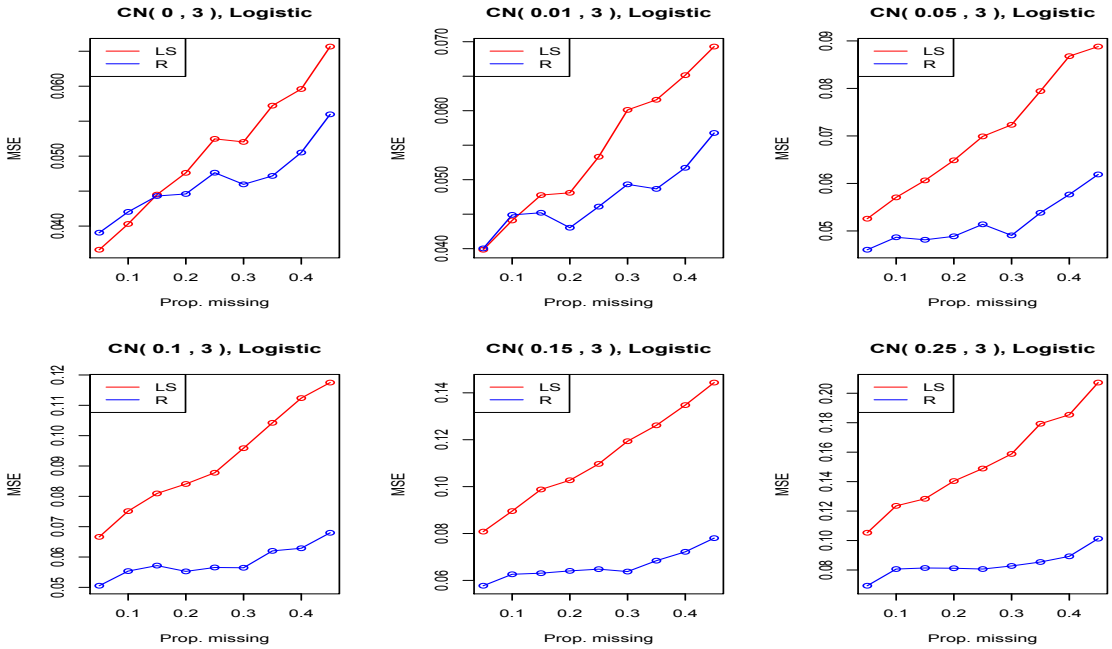


Figure 4.8: Scenario 1, Case 3: MSE vs Proportion of Missing Data for inverse probability (Log) under t error distribution

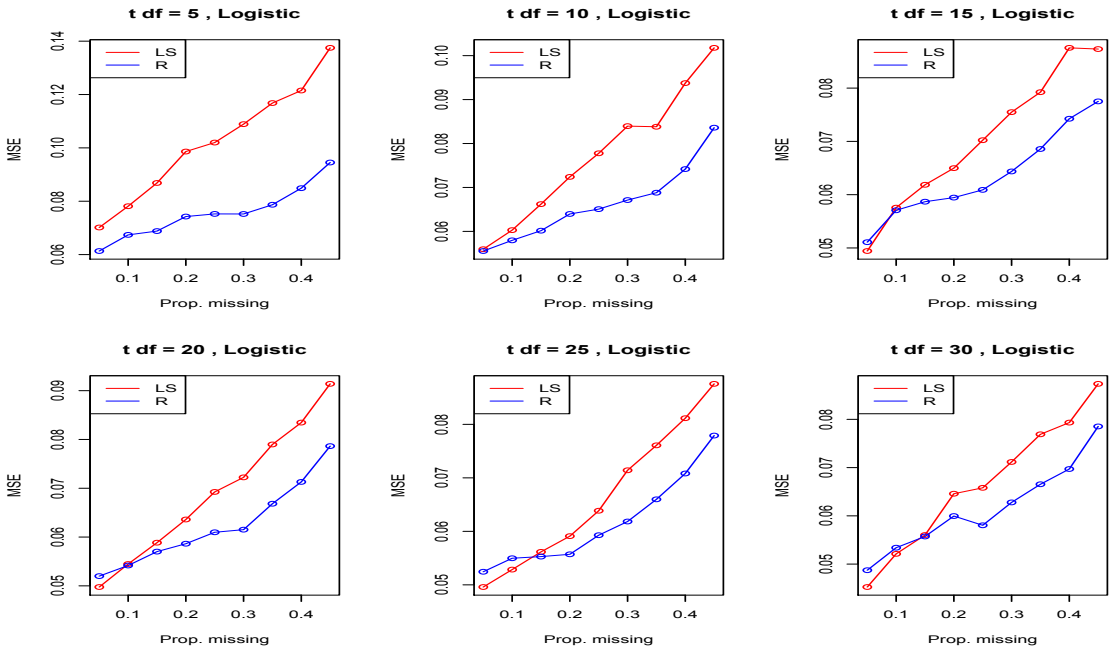


Figure 4.9: Scenario 2, Case 1: MSE vs Proportion of Contamination and t-df

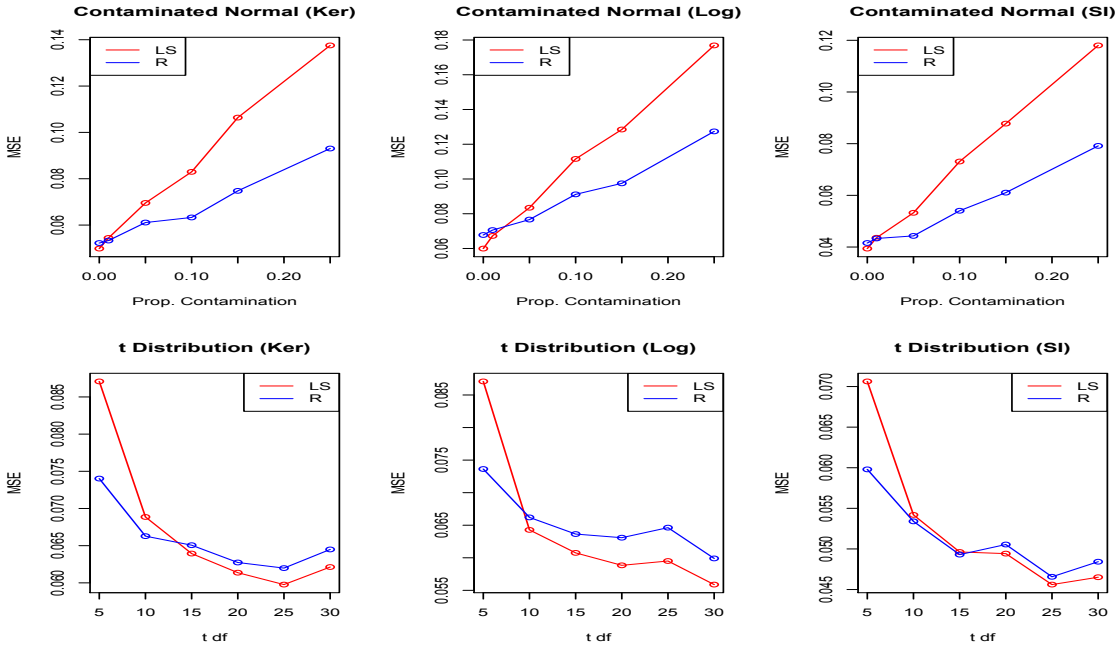


Figure 4.10: Scenario 2, Case 2: MSE vs Proportion of Contamination and t-df

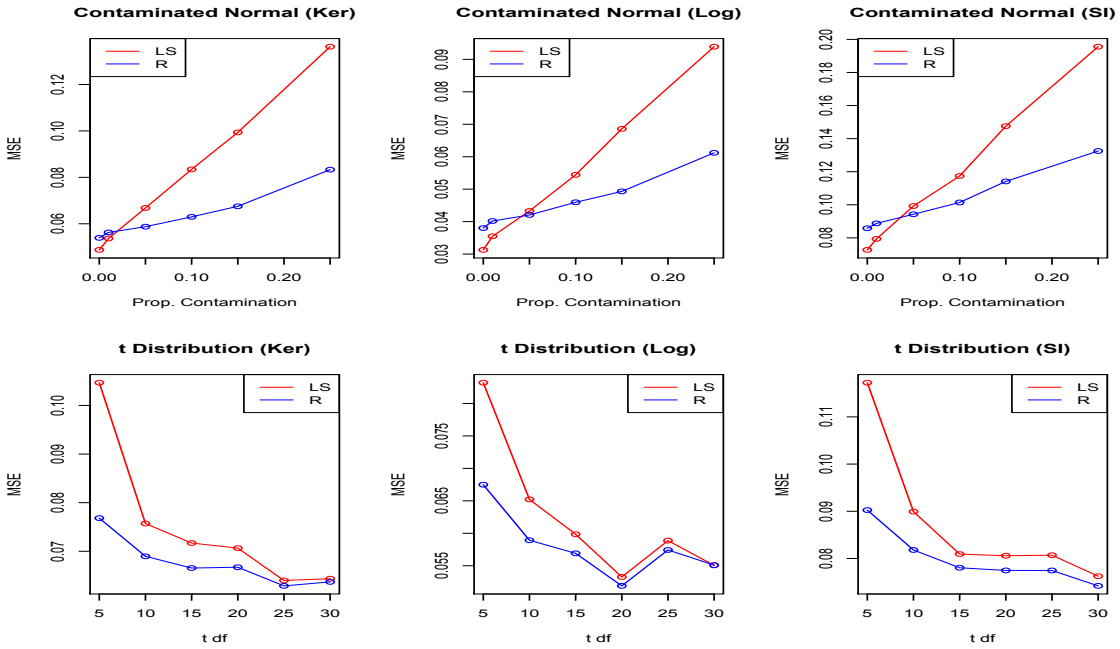


Figure 4.11: Scenario 2, Case 3: MSE vs Proportion of Missing Data for SI under contaminated normal error distribution

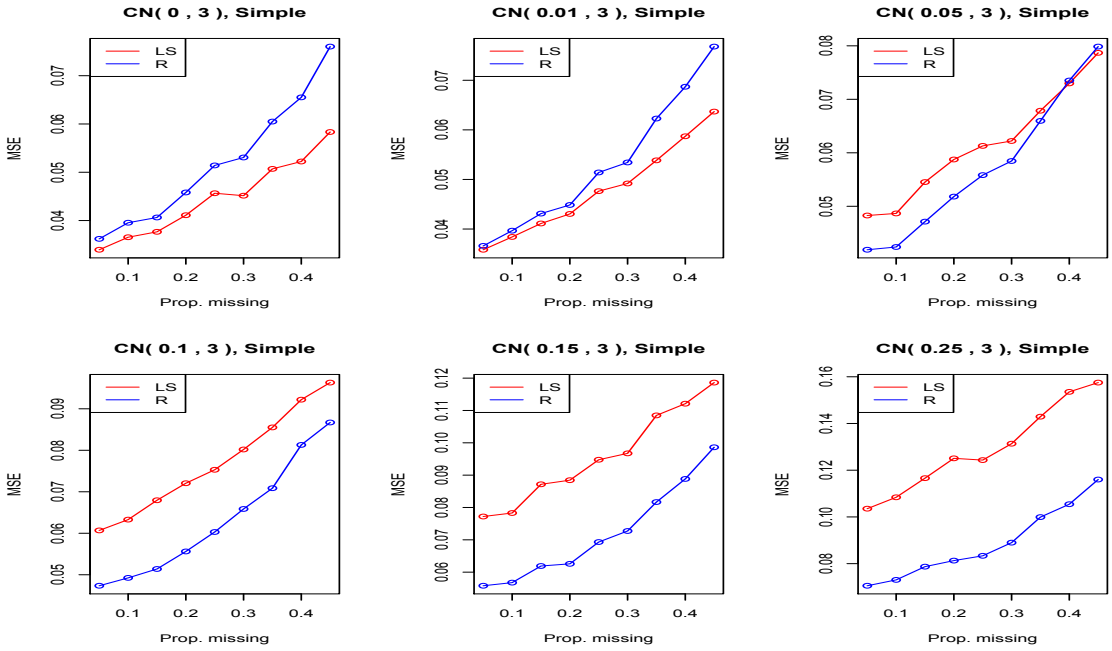


Figure 4.12: Scenario 2, Case 3: MSE vs Proportion of Missing Data for SI under t error distribution

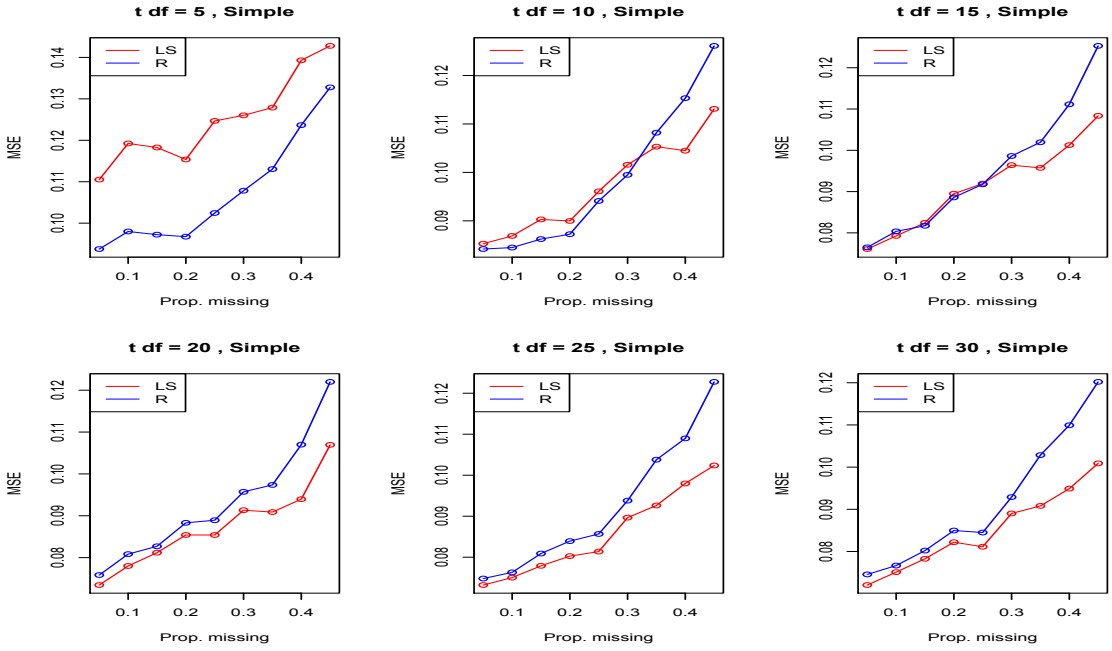


Figure 4.13: Scenario 2, Case 3: MSE vs Proportion of Missing Data for inverse probability (Ker) under contaminated normal error distribution

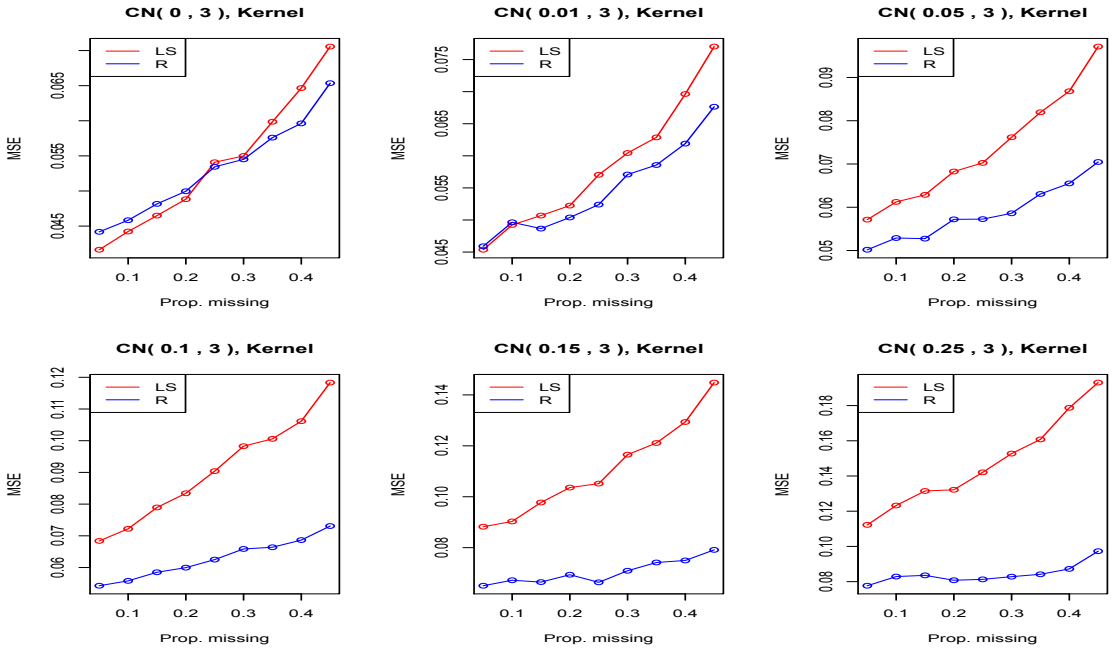


Figure 4.14: Scenario 2, Case 3: MSE vs Proportion of Missing Data for inverse probability (Ker) under t error distribution

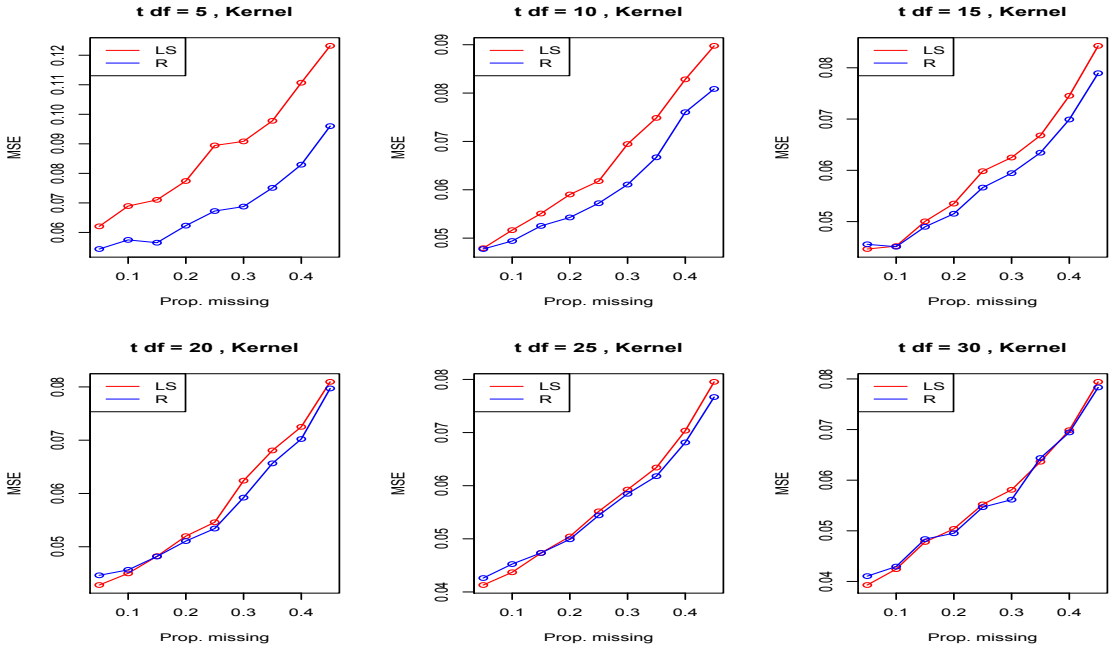


Figure 4.15: Scenario 2, Case 3: MSE vs Proportion of Missing Data for inverse probability (Log) under contaminated normal error distribution

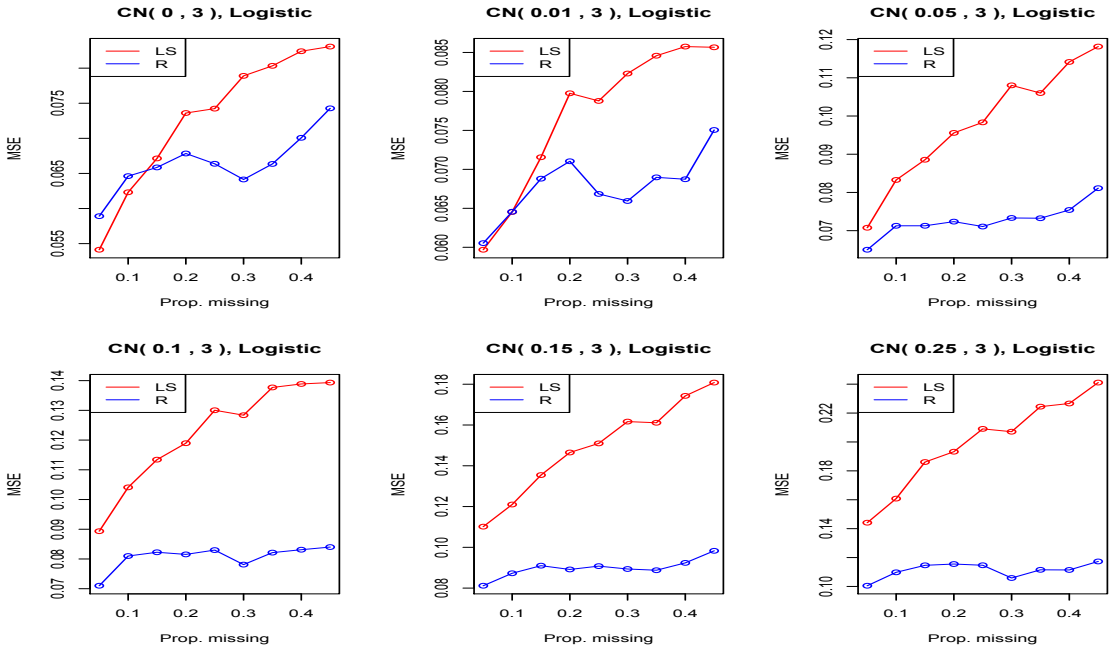
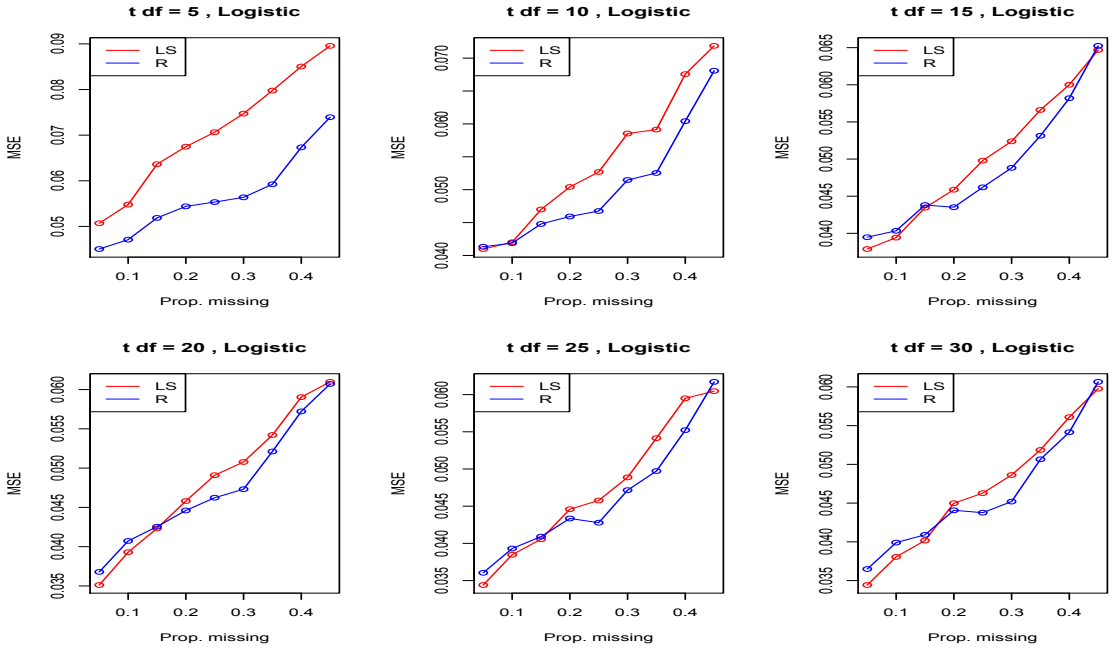


Figure 4.16: Scenario 2, Case 3: MSE vs Proportion of Missing Data for inverse probability (Log) under t error distribution



Scenario 1:

Case 1: For the case of the CN distribution, the R estimator becomes more and more superior to LS when the proportion of contamination increased. The same is true under the t distribution with the R becoming superior to LS as the degrees of freedom decrease (heavier tails). The MSEs are generally larger for the logistic case in the contaminated normal case but smaller in the t distribution case.

Case 2: Once again R is increasingly superior to LS as the contamination increases or as the tail-thickness of the distribution increases. For CN distribution, (SI) generally does worse than (Log) and (Ker) and for t distribution (Log) does worse than (Ker) and (SI).

Case 3: Under (SI), R performs comparably to LS under normality ($CN(0,3)$) or tails close to the tails of the normal distribution ($CN(.01,3)$ *and* t_{30}) but, as in the previous cases, does much better under heavy tails or larger contamination. What is additionally notable under (Ker) and (Log) is that R consistently outperforms LS when the proportion of missing data is large, even under normality.

Scenario 2:

Case 1: Similar to Scenario 1, Case 1.

Case 2: Similar to Scenario 1, Case 2.

Case 3: Similar patterns are observed as in Case 3 or Scenario 1. There are, however, some notable differences. Under Scenario 1, for the (SI) case, R and LS were comparable under normality. Under Scenario 2, LS is clearly superior to R. Under Scenario 1, for the (Ker) case, R does better than LS when the proportion of missing increases. This is no longer the case when the t degree of increases where R and LS are comparable. A similar observation can be made for the (Log) case.

Case 4: Similar to Case 2.

The first take-home message is that R performs better under heavy tailed error distributions and cases containing contaminations. It is generally comparable to LS under normal error. The second take-home message is that R with inverse probability imputation does better than its LS counterpart when the proportion of missing data is large. This makes R estimation extremely appealing for situations where we encounter high rates of missing information.

We finish this section by giving an alternative approach to defining rank estimators under responses missing at random. The theory follows from the above theorems in a (mostly) straightforward manner.

4.10 Empirical log-likelihood approach

Setting

$$\eta_{ij}(\beta) = \varphi\Big(\frac{R^l(\mathbf{v}_n^{ij}(\beta))}{n+1}\Big)\big(X_i - \tilde{g}_{1n}(T_i)\big)$$

for $j = 1, 2$. One can define the empirical log-likelihood function of β as follows

$$L_n^{ij}(\beta) = -2 \sup_{(p_1,\ldots,p_n)\in(0,1)^n} \Big\{ \sum_{i=1}^n \log(np_i) : \ \sum_{i=1}^n p_i = 1, \ \sum_{i=1}^n p_i\eta_{ij} = 0\Big\}, \ j = 1, 2.$$

Now take $\lambda \in \mathbb{R}^d$ for which the following equation holds:

$$\frac{1}{n}\sum_{i=1}^n \frac{\eta_{ij}(\beta)}{1 + \lambda^\tau \eta_{ij}(\beta)} = 0. \tag{4.12}$$

From this, $L_n^{ij}(\beta)$ can be rewritten as

$$L_n^{ij}(\beta) = 2\sum_{i=1}^n \log\big(1 + \lambda^\tau \eta_{ij}(\beta)\big). \tag{4.13}$$

Consider the matrix A_{nj} defined by $A_{nj} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n \eta_{ij}(\beta_0)\eta_{ij}^\tau(\beta_0)$. Clearly from the fact that $E(\varepsilon_i|Z_i) = 0$ and $E(\varepsilon_i^2|Z_i) < \infty$, $A_{nj} \to \mathbf{V}_j$ as $n \to \infty$. Based on Theorem 12, we have

$\max_{1\leq n} \|\eta_{ij}(\beta_0)\| = o_P(c_n)$ for some c_n. Also, φ being bounded, implies that there exists a positive constant C such that $c_n = C\alpha(n)$. Hence, one can obtain the asymptotic distribution of $L_n^{ij}(\beta)$ at the true parameter. This is given in the following theorem.

Theorem 16. *Under the assumptions* $(J_1) - (J_9)$*, we have*

$$L_n^{ij}(\beta_0) \xrightarrow{\mathcal{D}} \chi_p^2$$

The proof of this Theorem is obtained by putting together Theorem 12 and Slutsky's Lemma.

Remark 11. *Note that considering the right hand side of* (4.13) *as a function of* λ *and performing the Taylor expansion of order 1 at* $\eta_{ij}(\beta_0)$ *and using* (4.12)*, we obtain*

$$L_n^{ij}(\beta_0) = \frac{A_{nj}^{-1}}{n}\Big(\sum_{i=1}^{n}\eta_{ij}^{\tau}(\beta_0)\Big)\Big(\sum_{i=1}^{n}\eta_{ij}(\beta_0)\Big)(1+o_p(1)).$$

Setting

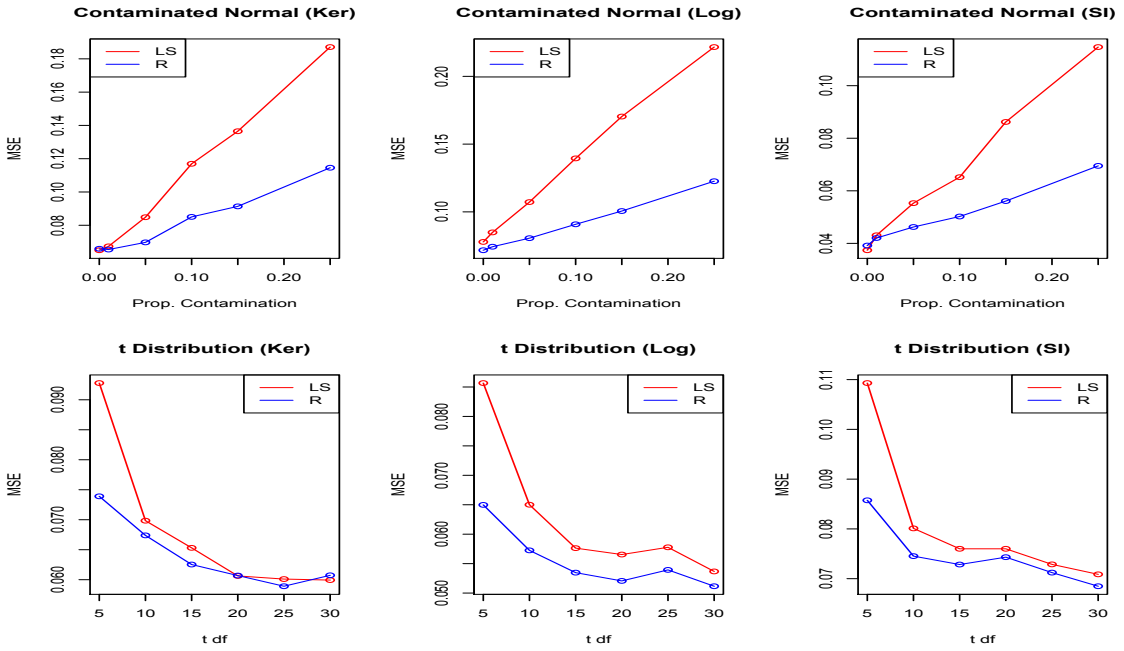
$$\mathcal{L}_n^{ij}(\beta_0) = \frac{A_{nj}^{-1}}{n}\Big(\sum_{i=1}^{n}\eta_{ij}^{\tau}(\beta_0)\Big)\Big(\sum_{i=1}^{n}\eta_{ij}(\beta_0)\Big) = nA_{nj}^{-1}\big(S_n^l(\beta_0)\big)^{\tau}S_n^l(\beta_0),$$

we see that for n *large enough,* $L_n^{ij}(\beta_0) \approx \mathcal{L}_n^{ij}(\beta_0)$*. Hence,* $\mathcal{L}_n^{ij}(\beta_0) \xrightarrow{\mathcal{D}} \chi_p^2$*. Now putting* $\beta_\varphi^* = \underset{\beta\in\mathcal{B}}{\text{Argmax}}\{-L_n^{ij}(\beta)\}$*, the maximum empirical likelihood estimator of* β_0*, we have following theorem.*

Theorem 17. *Under the assumptions* $(J_1) - (J_9)$*,*

$$\sqrt{n}(\beta_\varphi^* - \beta_0) \xrightarrow{\mathcal{D}} N(0, \gamma_\varphi^2\Sigma_j^{-1}),$$

Figure 4.17: Scenario 2, Case 4: MSE vs Proportion of Contamination and t-df



Bibliography

Abebe, A. and McKean, J. W. (2007). Highly efficient nonlinear regression based on the Wilcoxon norm. In Umbach, D., editor, *Festschrift in Honor of Mir Masoom Ali*, pages 340–357.

Adams, R. A. (1975). *Sobolev spaces.* Academic Press, New York-London.

Brezis, H. (1983). *Analyse fonctionnelle: Théorie et applications.* Masson, Paris.

Brunner, E. and Denker, M. (1994). Rank statistics under dependent observations and applications to factorial designs. *Journal of Statistical Planning and Inference*, 42(3):353 – 378.

Chang, W. H., McKean, J. W., Naranjo, J. D., and Sheather, S. J. (1999). High-breakdown rank regression. *J. Amer. Statist. Assoc.*, 94(445):205–219.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):pp. 1166–1181.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):pp. 81–87.

Coakley, C. W. and Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *J. Amer. Statist. Assoc.*, 88(423):872–880.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):pp. 1380–1403.

Gay, D. M. (1983). ALGORITHM 611 – subroutines for unconstrained minimization using a model/trust-region approach. *ACM Trans. Math. Software*, 9(4):503–524.

Gay, D. M. (1984). A trust-region approach to linearly constrained optimization. In Griffiths, D. F., editor, *Numerical Analysis. Proceedings, Dundee 1983*, pages 72–105. Springer-Verlag. Lecture Notes in Mathematics 1066.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986). Some new estimation methods for weighted regression when there are possible outliers. *Technometrics*, 28(3):219–230.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics*, 50(3):pp. 640–652.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. A roughness penalty approach.

Hájek, J. and Šidák, Z. (1967). *Theory of rank tests*. Academic Press, New York.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58(2):109–127.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393.

Haupt, H. and Oberhofer, W. (2009). On asymptotic normality in nonlinear regression. *Statist. Probab. Lett.*, 79(6):848–849.

Healy, M. and Westmacott, M. (1956). Missing values in experiments analysed on automatic computers. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 5(3):pp. 203–206.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2):pp. 244–248.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust nonparametric statistical methods*, volume 5 of *Kendall's Library of Statistics*. Edward Arnold, London.

Hettmansperger, T. P., McKean, J. W., and Sheather, S. J. (2000). Robust nonparametric methods. *J. Amer. Statist. Assoc.*, 95(452):1308–1312.

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.*, 23(1):92–119.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, 43:1449–1458.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40:633–643.

Jureckova, J. (1971). Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42(4):pp. 1328–1338.

Jurečková, J. (2008). Regression rank scores in nonlinear models. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. Collect.*, pages 173–183. Inst. Math. Statist., Beachwood, OH.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25(5):pp. 2084–2102.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.*, 77(379):595–604.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, 37(5A):2245–2277.

Naranjo, J. D. and Hettmansperger, T. P. (1994). Bounded influence rank regression. *J. Roy. Statist. Soc. Ser. B*, 56(1):209–220.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):pp. 90–120.

Peng, L. (2004). Empirical-likelihood-based confidence interval for the mean with a heavy-tailed distribution. *The Annals of Statistics*, 32(3):pp. 1192–1214.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & Probability Letters*, 4(4):203 – 208.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):pp. 846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):pp. 931–954.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflug, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, volume 8, pages 283–297. Reidel, Dordrechet.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

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

Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Comm. Statist. A—Theory Methods*, 12(10):1161–1179.

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

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):pp. 413–436.

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *Journal of the American Statistical Association*, 88(421):pp. 237–244.

Stromberg, A. J. (1995). Consistency of the least median of squares estimator in nonlinear regression. *Comm. Statist. Theory Methods*, 24(8):1971–1984.

Stromberg, A. J. and Ruppert, D. (1992). Breakdown in nonlinear regression. *J. Amer. Statist. Assoc.*, 87(420):991–997.

Sun, Z., Wang, Q., and Dai, P. (2009). Model checking for partially linear models with missing responses at random. *J. Multivar. Anal.*, 100(4):636–651.

Tableman, M. (1990). Bounded-influence rank regression: a one-step estimator based on Wilcoxon scores. *J. Amer. Statist. Assoc.*, 85(410):508–513.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Van Zwet, W. R. (1980). A strong law for linear functions of order statistics. *The Annals of Probability*, 8(5):pp. 986–990.

Wang, C. Y., Wang, S., Zhao, L.-P., and Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92(438):pp. 512–525.

Wang, J. (1996). Asymptotics of least-squares estimators for constrained nonlinear regression. *Ann. Statist.*, 24(3):1316–1326.

Wang, J. D. (1995). Asymptotic normality of L_1-estimators in nonlinear regression. *J. Multivariate Anal.*, 54(2):227–238.

Wang, Q., Linton, O., and HŁrdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 99(466):pp. 334–345.

Wang, Q. and Rao, J. N. K. (2002a). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, 30(3):pp. 896–924.

Wang, Q. and Rao, J. N. K. (2002b). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.*, 30(3):896–924. Dedicated to the memory of Lucien Le Cam.

Wang, Q. and Sun, Z. (2007). Estimation in partially linear models with missing responses at random. *Journal of Multivariate Analysis*, 98(7):1470 – 1493.

Wiens, D. and Zhou, J. (1994). Bounded-influence rank estimation in the linear model. *Canad. J. Statist.*, 22(2):233–245.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.*, 9(3):501–513.

Xue, L. (2009). Empirical likelihood confidence intervals for response mean with data missing at random. *Scandinavian Journal of Statistics*, 36(4):671–685.

Xue, L. and Zhu, L. (2007a). Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, 102(478):642–654.

Xue, L. and Zhu, L. (2007b). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, 94(4):921–937.

Xue, L.-G. and Zhu, L. (2006). Empirical likelihood for single-index models. *Journal of Multivariate Analysis*, 97(6):1295 – 1312.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, 15(2):642–656.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50(3):pp. 689–699.

Zhao, L. P., Lipsitz, S., and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics*, 52(4):pp. 1165–1182.

Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):pp. 1053–1068.