12-2002

On Rank-Based Considerations for Generalized Linear Models and Generalized Estimating Equation Models

Diana R. Cucos
Western Michigan University

ON RANK-BASED CONSIDERATIONS FOR GENERALIZED LINEAR MODELS AND GENERALIZED ESTIMATING EQUATION MODELS

by

Diana R. Cucos

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

Western Michigan University
Kalamazoo, Michigan
December 2002

ON RANK-BASED CONSIDERATIONS FOR GENERALIZED LINEAR MODELS AND GENERALIZED ESTIMATING EQUATION MODELS

Diana R. Cucos, Ph.D.

Western Michigan University, 2002

This study discusses rank-based robust methods for estimation of parameters and hypotheses testing in the generalized linear models (GLM) and generalized estimating equations (GEE) setting. The robust estimates are obtained by minimizing a Wilcoxon drop in dispersion function for linear or nonlinear regression models. In addition, diagnostic tools for outliers and influential observations are being developed. These models are generalizations of linear and nonlinear models. They allow for both nonlinear mean functions and heteroscedasticity of their random errors. This makes them quite useful in practice.

Rank-based inference has been developed for linear models over the last thirty years. This inference is both robust and highly efficient and it can be extended to estimates which have high breakdown. It has recently been extended to nonlinear models. In this work, we extend this inference to GLM and GEE models.

The robust estimates of the mean function are obtained by minimizing a norm based on Wilcoxon scores in much the same way least squares type estimates are obtained by minimizing the Euclidian norm. For the heteroscedasticity

problem where the errors are independent but have non-constant variances, we show that these robust estimates retain their consistency and asymptotic normality provided scale is consistently estimated. We further develop asymptotic theory for robust testing based on both Wald type tests and drop in dispersion tests. In addition, diagnostic tools for outliers and influential observations are developed. We discuss extensions to high-breakdown estimates. We discuss a robust estimate of the variance-covariance matrix for the auto-regressive structure, used for the GEE models.

Examples and simulation studies illustrate the robustness of the procedure and its superiority against the classical statistical techniques currently used. Data for the examples include a multiple sclerosis longitudinal trial and a cholesterol data from randomly selected individuals from the Framingham study.

UMI Number: 3077375

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGMENTS

I will never find enough words to express my gratitude to my dissertation advisor, Dr. Joseph W. McKean. His continued guidance, his endless supply of advice and ideas made this research possible. Regardless of where the career path will take me, I will always hold close his advice and when in doubt about methods or ethics, I will always go back and ask myself whether he would approve of my choices.

I would like to thank the members of my committee, Dr. Joseph W. McKean, Dr. Michael Stoline, Dr. Joshua Naranjo, Dr. Jerry Sievers, and Dr. Tom Vidmar, for reviewing my work and providing their feedback. I would also like to thank the faculty of the Department of Statistics. I enjoyed each and every one of the classes taken during my time as a doctoral student, and their knowledge and passion laid the foundation for this research.

I am forever grateful to my parents for their support and love, and to my son and husband for their understanding and patience.

Diana R. Cucos

TABLE OF CONTENTS

Table of Contents—Continued

Table of Contents—Continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

1.1.1 Generalized Linear Models

Nelder and Wedderburn (1972) were the first to unify regression and linear models into the generalized linear model (GLM) and to propose a computational method for finding the Maximum Likelihood Estimators (MLE). GLM has been used widely in biomedical research for exploring relationships and for estimating the effect of a set of covariates on an outcome variable or disease status. In such models the mean response is related to a vector of regressor variables through a link function. Common link functions are the identity, logit, probit, power, log, complementary log-log and other domain-specific functions (e.g. Michaelis-Menten relation for pharmacokinetic data). However, generalized linear models (McCullagh and Nelder (1989)) require the observations to be independent, and to follow an exponential family distribution (often normal, gamma, Poisson or binomial). This relationship may be expressed as

$$f(y_i) = \exp\{\frac{y_i\theta_i - a(\theta_i)}{\phi_i}\}h(y_i, \phi_i), \quad g(\mu(\theta_i)) = \beta X_i^T, \tag{1.1}$$

with θ_i, ϕ_i unknown and $\phi_i > 0$, $\phi_i = \frac{\phi}{t_i}$, $\phi > 0$ unknown, and $t_i > 0$ known.

Some of the most interesting studies, though, involve repeated measurements taken on same subject, over time. Successive measurements are usually dependent. Other terms used for repeated measures are longitudinal data (used to imply that the patients are followed over time), panel study and growth data (subjects measured at a common set of ages). For example, in a clinical study, a large number of individuals may be examined for a successive number of months to determine their reaction to a new asthma medication, and their pulmonary function measurements (continuous response) and other possible covariates (both static, eg. gender, race, age, severity of disease, comorbidities; and time-varying, e.g. weight, height) recorded. The measurements (over time) for a particular patient are almost never independent.

1.1.2 Extensions to Generalized Linear Models

Jorgensen (1983) proposed an extension to GLM which allowed for correlated errors and nonlinear hypotheses. The extended class also admitted distributions that are not members of the exponential family. He considered examples which included multivariate normal, log gamma , hyperbolic and inverse Gaussian distributions.

Since then, GLM has been extended in three directions:

(1) Conditional (or subject-specific) models (Rosner (1985)),

(2) Random effects models (mixed models), and

(3) Marginal (or population-averaged) models.

Because the random effects are not observed data, in the likelihood form of the conditional models, they are integrated out. This integration does not have a closed form solution for non-normally distributed data. The popular numerical approximation methods (implemented by SAS) do not work well for longitudinal data with high correlation (Breslow and Lin (1995)).

1.1.3 Generalized Estimating Equations

Liang and Zeger (Liang and Zeger (1986), Zeger and Liang (1986)) extended GLM to correlated data using a marginal model, called Generalized Estimating Equations (GEE). If number of time points is one, then their method reduces to the GLM procedure. The GEE approach has its roots in the quasi-likelihood methods introduced by Wedderburn (1974) and developed and extended by McCullagh and Nelder (1989). Standard maximum-likelihood analysis required the specification of the full conditional distribution of the dependent variable (most commonly, assuming it to be normally distributed) and estimated parameters by solving the score equation (the derivative of the likelihood function with respect to the parameters set equal to zero). Instead, the quasi-likelihood methods require only that the relationship between the expected value of the dependent variable (Y) and the covariates (X) is known. The parameters are estimated by solving a quasi-score equation, given by

$$\text{Score:} \sum_{i=1}^{K} X_i^T \boldsymbol{\delta}_i (Y_i - a'_i(\boldsymbol{\theta})) = 0,$$

$$\text{Quasiscore:} \sum_{i=1}^{K} D_i^T V_i^{-1} (Y_i - a'_i(\boldsymbol{\theta})) = 0,$$

$\boldsymbol{\delta}_i = \text{diag}(\frac{\partial \theta_{it}}{\partial \eta_{it}})$, $\eta_{it} = \mathbf{x}_{it}^{\mathbf{T}} \beta$, $D_i = A_i \boldsymbol{\delta}_i X_i$, $V_i = \frac{A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}}{\phi}$, $R(\boldsymbol{\alpha})$ known matrix for $\boldsymbol{\alpha} > 0$, $A_i = \text{diag}(a''(\theta_{it}))$.

These marginal models are useful for any situation in which the emphasis is on understanding the relationship between the regressor variables and the mean response variable, where data are correlated but the correlations are not the main focus of the analysis. A detailed discussion of Liang and Zeger model and results is presented in Section 2.2.

GEE's are estimable with a host of software packages. SAS's Proc Genmod estimates GEE for normal, binomial, Poisson and gamma families. Available correlation structures are independent, exchangeable, AR(1), m-dependent and pairwise. S-plus has a user-written GEE module (formerly OSWALD). Available links are identity, log, logit, probit, reciprocal and complimentary log-log. Another readily available computer package is BMDP.

GEE's are used in practice for binary response variables, for event counts, for ordered or unordered polychotomous, and for continuous responses. The present study focuses on continuous responses.

1.1.4 Correlation Structure

To increase the efficiency of the estimators, GEE's take into consideration the fact that the correlation structure is probably not that of independence. Liang and Zeger (1986) assume the *working* correlation structure is known (from previous experience, or the physics of the phenomena being studied). If the *working* correlation matrix is correctly specified, with its structure constant over clusters, then the GEE estimator is consistent and asymptotically normal. Liang and Zeger proposed a *robust* (or *empirically-corrected*) estimate of the variance-covariance matrix, which guarantees the consistency of the estimator even under misspecification of the correlation matrix. This estimator is also discussed in Section 2.2.

The correlation structures often used in practice are the following.

(1) Independence - This assumes no intra-cluster correlation. The estimator is the analogue of *pooled* estimators.

(2) Exchangeable - The correlation is assumed to be equal across all observations within a cluster. In this case only one parameter needs to be estimated. The model is analogous to the random-effects model.

(3) Unstructured - This places no constraints on the correlation structure. Thus all pairwise correlations need to be estimated.

(4) Autoregressive - The correlation over time is modeled as an exponential

function of the lag length.

The AR(1) is the correlation structure used most of the time, especially if data are equally spaced. Even for unequally spaced data, a continuous time AR(1) correlation structure has been proposed by Jones and Boadi-Boateng (1991). For an AR(1), one parameter needs to be estimated.

In addition, the researcher may explicitly specify the correlation structure. Diggle, Liang and Zeger (1995) recommend that:

> When the regression coefficients are the scientific focus ... one should invest the lions share of time in modeling the mean structure, while using a reasonable approximation to the covariance. The robustness of the inferences about β can be checked by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance model may be necessary.

Miller (1993) used a *working* correlation structure based on the inverse of Fisher's z transformation. But estimation in this case (see Prentice and Zhao (1991)) requires another *working* correlation matrix for the sum of squares and sum of products variables.

In general, the consistency of the estimator of θ may be established even when the 'working' correlation matrix is incorrectly specified; hence, the assumption of experience-dictated correlation matrix will be used throughout this research. That is, we will assume the structure of the correlation model, but the correlation will have unknown parameters which we will estimate in Section 3.5.

Liang and Zeger (1986) assumed that the longitudinal correlation structure remains the same for all individuals. This approach reduces the number of correlation parameters to be estimated and it is widely applicable in practice. Least Squares (LS) and moment methods have been used in order to obtain consistent estimators of the correlations. In the present research, the correlation parameters (same across subjects) will be estimated using robust methods.

An alternative developed by Liang, Zeger and Qaqish (1992) allows for the joint estimation of the regression coefficients (including the intercept) and it is more efficient than the original estimation method (often referred to as GEE_1). Unlike GEE, this technique (GEE_2), obtains estimates using the first two empirical moments instead of just the first empirical moment. However, the mean and the correlation structure need to be correctly specified, and this estimator is even more sensitive to departures from the true covariance structure. Hence, it is even less robust than the traditional GEE.

1.1.5 Diagnostics for GEE

Although consistent, GEE estimators are not efficient (Crowder (1995), Sutradhar and Das (1999)) and they are not robust in the presence of outlying observations. At the present time, standard residual diagnostics are employed, alongside plots of residuals stratified by time or by subject. Goodness-of-fit is assessed using Schwartz's Bayesian Criterion and the Akaike's Information Criterion, both functions of the likelihood. However, since the GEE residuals are correlated, summary goodness-of-fit statistics may not be appropriate. Zheng (2000) extended four goodness-of-fit measures of GLM to GEE models: proportional reduction in entropy, percent reduction in deviance, concordance correlation coefficient and concordance index.

1.1.6 Robust Methods

Considerable effort was put into robustifying methods for longitudinal data. Morgenthaler (1992) replaced the L_2 norm by the L_1 norm in the derivation of the quasi-likelihoods. However, the extension leads to biased estimating equations. In order to obtain consistent estimators, the underlying distribution needs to be known. The mean and the covariance structure alone do not suffice anymore.

Chi and Reisel (1989) proposed a mixed linear model containing fixed regression parameters, random effects across individuals and autocorrelation in the within-individual errors. Gill (2000) robustified the log-likelihood for Chi's model, using Huber's ρ function. Details of this estimator are provided in Section 2.3.1.

Hu and Lachin (2001) developed a robust alternative to the GEE's, called truncated robust estimating equations, not sensitive to heavy-tailed distributions, contaminated distributions or extreme values. Their estimator is based on truncated estimating functions (namely, on Huber's M estimator). The robust estimator is more efficient for non-normal data, and it is not as sensitive to departures from the true correlation structure. As pointed by the authors, the danger in using M-estimators is to over-truncate the tail values, which may contain some useful information. The details of this estimator are presented in Section 2.3.2.

Robust alternatives were studied extensively for non-continuous data: M estimation, applied to binomial and poisson models (Cantoni and Ronchetti (2001)); M estimation applied to logistic regression (Adimari and Ventura (2002)); resistant generalized estimating equations that include weights in the estimating equations to downweight influential observations (weights are computed according to the observation leverage (Mallows class downweights), or residual (Schweppe class downweights) (Preisser and Qaqish (1999)).

The present research will investigate rank-based robustness, as an extension to Generalized Estimating Equations. The theory will follow as an extension from the theory for rank-based analysis of GLM, presented by Hettmansperger and McKean (1998). The complete inference in Hettmansperger and McKean (1998) holds for the general rank scores. For the present research, however, we have chosen the Wilcoxon scores.

1.2 Motivating Example

Gill (2000) compared his robust estimators with two previously published GEE estimators for a longitudinal clinical trial on multiple sclerosis patients (see details in Section 6.1). The data have been analyzed by Petkau and White (1995) and D'yachkova et al. (1997) using GEE's, although the data (or its transformation) are not normally distributed.

The study was a 3-year double-blind placebo-controlled randomized trial of interferon beta-1b given at two dosing schedules: 0.05 mg every other day, and 0.25 mg every other day. Each patient had a baseline cranial magnetic resonance imaging (MRI), and this was repeated yearly.

As part of the clinical trial, a cohort of 52 patients had head MRIs repeated at approximately 6-week intervals for two years. Data from 49 patients were used in the analysis. The patient *burden of disease* was used as an indicator of the severity of the disease at the time of the scan. Graphical analyses of the data revealed strong skewness in the burden measurements. Petkau and White (1995) suggested a log transformation to normalize the data. However, as shown in Figure 1, the transformed data has very long tails, with outlying observations mainly in the placebo and low-dose group (high-dose group is close to normal).

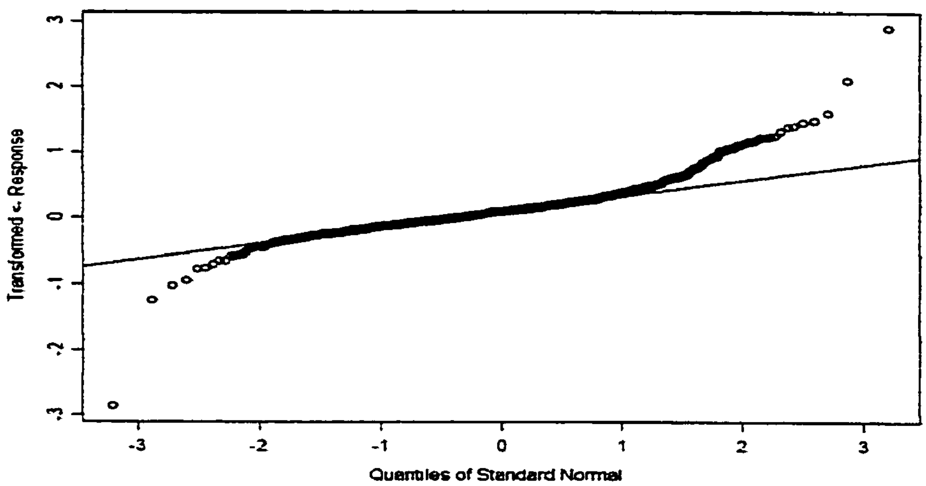

Figure 1. Q-Q Plot of Log-Transformed Multiple Sclerosis Data

Gill's estimators of slope parameters (see Table 1) show a significant difference in the estimators for the placebo and low-dose group, with the high-dose groups estimators being practically equal to their GEE counterparts. The standard errors of his robust estimators are lower than the GEE standard errors.

Table 1

LS Estimates Versus Gill's Robust Estimates

	Placebo	Low Dose	High Dose
MLE	0.036	0.032	-0.095
Gill's Robust	0.021	0.038	-0.094

Since Gill's estimators are a M-robust version for the mixed linear models, it would be of interest to compare the LS results with those of a direct rank-based robust extension of GEE.

CHAPTER II

CLASSICAL GEE THEORY AND EXTENSIONS

2.1 Notation

We shall use the notation in Liang and Zeger (1986).

Let $\mathbf{Y}_i^{n\times 1} = (y_{i1}, \ldots, y_{in})^T$ be the vector of outcome values and $X_i^{n\times p} = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in})^T$ be the matrix of covariates for the ith subject, $i = 1, \ldots, K$, $t = 1, \ldots, n$.

Assume that the marginal distribution of y_{it}, $i = 1, \ldots, K$, $t = 1, \ldots, n$ is :

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi], \tag{2.1}$$

where $a(\cdot)$ and $b(\cdot)$ are continuous functions, $\boldsymbol{\theta}$ and ϕ are model parameters, and we are fitting the model

$$E(y_{it}) = g(\theta_{it}), \quad \theta_{it} = h(\eta_{it}), \quad \eta_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta}, \tag{2.2}$$

where $g(\cdot)$ and $h(\cdot)$ are continuous functions and $\boldsymbol{\beta}^{p\times 1}$ is a vector of unknown coefficients.

Equivalently, the mean for each subject may be expressed as

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = [g(h(\mathbf{x}_{i1}^T(\boldsymbol{\beta})), \ldots, g(h(\mathbf{x}_{in}^T\boldsymbol{\beta}))]^T.$$

Using a moment-generating-function argument, it follows that the first two moments of y_{it} are given by

$$E(y_{it}) \equiv \boldsymbol{\mu}_i(\boldsymbol{\beta}) = a'(\theta_{it}),$$

and

$$\mathrm{var}(y_{it}) = \frac{a''(\theta_{it})}{\phi} = \sigma_{it}^2.$$

Let

$$\Delta_i^{n \times n} = \mathrm{diag}(\frac{\partial \theta_{it}}{\partial \eta_{it}}) \quad \text{and} \quad A_i = \mathrm{diag}(a''(\theta_{it})).$$

Furthermore, let $R(\boldsymbol{\alpha})$ be a $n \times n$ symmetric matrix with $\boldsymbol{\alpha}^{s \times 1}$, $s > 0$, a vector which fully characterizes $R(\boldsymbol{\alpha})$. $R(\boldsymbol{\alpha})$ is called the *working* correlation matrix.

2.2 Least Squares Theory

Under the independence assumption, the Fisher score equation from a likelihood analysis has the form

$$U_I(\boldsymbol{\beta}) = \sum_{i=1}^{K} X_i^T \Delta_i (\mathbf{Y}_i - \mathbf{a}_i'(\boldsymbol{\theta})) = 0 \tag{2.3}$$

The estimator $\hat{\boldsymbol{\beta}}_I$ is the solution to the score equation.

Liang and Zeger (1986) proposed a class of estimating equations which take the correlation $R(\boldsymbol{\alpha})$ into account.

Let V_i be an estimator for $\text{cov}(\mathbf{Y}_i)$, based on $R(\boldsymbol{\alpha})$, and given by

$$V_i = \frac{A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}}{\phi}.$$

Then the General Estimating Equations are defined to be

$$\sum_{i=1}^{K} D_i^T V_i^{-1} (\mathbf{Y}_i - \mathbf{a}_i'(\boldsymbol{\theta})) = 0, \quad D_i = A_i \Delta_i X_i. \tag{2.4}$$

The estimator $\hat{\boldsymbol{\beta}}_G$ is the solution to the quasi-score equation, and it is called *GEE estimator*.

Note that (2.4) reduces to (2.3) if $R(\boldsymbol{\alpha})$ is the independence (identity) matrix.

Equation (2.4) is a function of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, but can be re-expressed as a function of $\boldsymbol{\beta}$ alone by replacing $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$, where

$$K^{\frac{1}{2}}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{\mathcal{P}} 0;$$

that is, $\hat{\boldsymbol{\alpha}}$ is $\sqrt{K}$ consistent when $\boldsymbol{\beta}$ and ϕ are known.

Provided that the mean is correctly specified, and regardless of the correct specification of $R(\boldsymbol{\alpha})$, the following two theorems hold; see Liang and Zeger (1986).

Theorem 2.2.1. *The estimator $\hat{\boldsymbol{\beta}}_G$ of $\boldsymbol{\beta}$ is consistent.*

The following theorem summarizes the asymptotic properties of $\hat{\beta}_G$.

Theorem 2.2.2. *Under mild regularity conditions and given that:*

(i) $\hat{\boldsymbol{\alpha}}$ is $K^{\frac{1}{2}}$-consistent given $\boldsymbol{\beta}$ and ϕ;

(ii) $\hat{\phi}$ is $K^{\frac{1}{2}}$-consistent given $\boldsymbol{\beta}$; and

(iii) $|\frac{\partial \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\phi)}{\partial \phi}| \le H(Y,\boldsymbol{\beta})$, where $H(Y,\boldsymbol{\beta}) \xrightarrow{\mathcal{P}} 0$,

then

$$K^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(0, V_G), \tag{2.5}$$

where

$$V_G = \lim_{K\to\infty} K\left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1}\left\{\sum_{i=1}^{K} D_i^T V_i^{-1} cov(\mathbf{Y}_i) V_i^{-1} D_i\right\}\left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1}.$$

Note that, similar to the quasi-likelihood theory, V_G does not depend on the choice of estimator for $\boldsymbol{\alpha}$ and ϕ. Thus, even when the exact nature of the dependence is not known, GEE offers asymptotically unbiased estimators.

To compute $\hat{\boldsymbol{\beta}}_G$, Liang and Zeger used the Gauss-Newton method: modified Fisher scoring for $\boldsymbol{\beta}$, calculate standardized residuals and then use moment estimation to obtain consistent estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

At each iterative step they compute

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} - \{\sum_{i=1}^{K} D_i^T(\hat{\boldsymbol{\beta}}^{(j)})\tilde{V}_i^{-1}(\hat{\boldsymbol{\beta}}^{(j)})D_j(\hat{\boldsymbol{\beta}}^{(j)})\}^{-1}\{\sum_{i=1}^{K} D_i^T(\hat{\boldsymbol{\beta}}^{(j)})\tilde{V}_i^{-1}(\hat{\boldsymbol{\beta}}^{(j)})\mathbf{S}_i(\hat{\boldsymbol{\beta}}^{(j)})\}, \tag{2.6}$$

where $\mathbf{S}_i = \mathbf{Y}_i - \mathbf{a}_i'(\boldsymbol{\theta})$ and $\tilde{V}_i(\boldsymbol{\beta}) = V_i[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}]$.

Note that this iterative step is the equivalent to performing an iteratively reweighted LS estimation of the linear regression of $Z = D\boldsymbol{\beta} - S$ on D with weight $\tilde{V}^{-1}$, where $\tilde{V}^{nK \times nK}$ is block diagonal with $\tilde{V}_i$ as diagonal elements.

In Zeger and Liang (1986) the following robust variance estimate, consistent even when V_i is incorrectly specified, was proposed:

$$V_{\hat{\boldsymbol{\beta}}} = M_0^{-1} M_1 M_0^{-1}, \quad \text{where}$$
$$M_0 = \sum_{i=1}^{K} D_i^T(\hat{\boldsymbol{\beta}})\hat{V}_i^{-1} D_i(\hat{\boldsymbol{\beta}}), \quad \text{and}$$
$$M_1 = \sum_{i=1}^{K} D_i^T(\hat{\boldsymbol{\beta}})\hat{V}_i^{-1}\mathbf{S}_i\mathbf{S}_i^T\hat{V}_i^{-1} D_i(\hat{\boldsymbol{\beta}}).$$

This estimator, called *sandwich variance estimator*, is frequently used in practice.

The scale parameter ϕ is estimated using the standardized residuals at a given iteration (i.e. generalized form of Pearson's statistic).

$$\hat{\phi}^{-1} = K \sum_{i=1}^{n} \frac{\hat{r}_{it}^2}{N-p}, \quad N = nK, \tag{2.7}$$

where $\hat{r}_{it}$ is the estimated standard residual

$$\hat{r}_{it} = \frac{y_{it} - a'(\hat{\theta}_{it})}{[a''(\hat{\theta}_{it})]^{\frac{1}{2}}}.$$

The correlation parameter $\boldsymbol{\alpha}$ is estimated as a function of

$$\hat{R}_{uv} = \sum_{i=1}^{K} \frac{\hat{r}_{iu}\hat{r}_{iv}}{N-p},$$

The form of the function is dictated by the choice of $R(\boldsymbol{\alpha})$.

In particular, for $AR(1)$, $\boldsymbol{\alpha}$ is estimated by the slope from the regression of $\log(\hat{r}_{it}\hat{r}_{it'})$ on $\log(|t-t'|)$.

2.3 Robust Estimates

2.3.1 Robustness in Mixed Linear Model setting

Gill (2000) proposed a robust procedure for estimating parameters of a mixed linear model, applied to longitudinal data.

Given K subjects, each observed on n time points, we let y_{it} be the continuous random variable corresponding to the response observed on the ith subject at time t. Assume the linear model to be

$$y_{it} = \mu_{it} + \epsilon_{it},$$

where μ_{it} is comprised of the general mean, time-dependent covariates, treatment effect and subject effect, and ϵ_{it} is the within-subject error. Errors from different subjects are assumed to be uncorrelated.

Hence,

$$\boldsymbol{\mu}_i = X_i\boldsymbol{\beta} + C_k\boldsymbol{\tau}_k,$$

C_k being the design matrix for the random effects.

Correlation among within-subject errors was assumed to follow an $AR(1)$ structure.

Let $\text{cov}(\mathbf{y}_i) = \sigma^2\Sigma$, where Σ is a function of an unknown parameter vector, $\boldsymbol{\alpha}$.

Following maximum-likelihood estimation, the log-likelihood for the generalized linear mixed models is:

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}|Y) = \text{constant} - \frac{1}{2}M\log(\sigma^2) - \frac{K}{2}\log|\Sigma| - \sum_{i=1}^{K}\frac{1}{2}\mathbf{r}'_i\mathbf{r}_i, \qquad (2.8)$$

where $\mathbf{r}_i = \sigma^{-1}\Sigma^{-\frac{1}{2}}(\mathbf{y}_i - X_i\boldsymbol{\beta})$.

Gill robustified this equation by replacing half of the sum of squares $\frac{1}{2}\mathbf{r}'_i\mathbf{r}_i$ by a robust function, namely Huber's ρ, function given by

$$\rho(a) = \begin{cases} \frac{1}{2}a^2 & \text{if} \quad |a| \leq c, \\ c|a| - \frac{1}{2}c^2 & \text{if} \quad |a| > c \end{cases}$$

where $c > 0$ is a fixed constant.

It is well known that the influence function for this function is bounded:

$$\psi(a) = \begin{cases} a & \text{if} \quad |a| \leq c, \\ c\text{sign}(a) & \text{if} \quad |a| > c. \end{cases}$$

Hence, Gill's robust version of the log-likelihood, to be maximized, is given by

$$\eta(\boldsymbol{\beta}, \boldsymbol{\alpha}|Y) = \text{constant} - \frac{1}{2}M\log(\sigma^2) - \frac{K}{2}k_1\log|\Sigma| - \sum_{i=1}^{K}\sum_{t=1}^{n}\rho(r_{it}),$$

where $k_1 = E(r\psi(r))$ is Huggins' consistency correction factor.

The parameter $\boldsymbol{\beta}$ is estimated using score equations. The iteration for the scoring procedure is given by

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \hat{\boldsymbol{\beta}}^{(h)} + (\hat{H}^{(h)}_{\beta\beta'})^{-1}\frac{\partial\eta^{(h)}}{\partial\boldsymbol{\beta}},$$

with $H_{\beta\beta'} = \nu\sigma^{-2}\sum_{i=1}^{K} X_i^T\Sigma^{-1}X_i$, the Hessian matrix for $\boldsymbol{\beta}$, $\nu = \Pr(|r| \leq c)$.

It has been shown that the asymptotic distribution of this estimator is

$$N_p(\boldsymbol{\beta}, \frac{E[\psi^2(r)]}{(E[\psi'(r)])^2}\hat{\sigma}^2(\sum_{i=1}^{K} X_i'\hat{\Sigma}^{-1}X_i)^{-1}). \qquad (2.9)$$

Note that if $c = \infty$, $\hat{\boldsymbol{\beta}}$ and its distribution reduce to the classical ML forms (under ML normality assumption, $\text{var}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{k} X_i'\hat{\Sigma}^{-1}X_i)^{-1}$.

To estimate the covariance parameters, $\boldsymbol{\alpha}$, in a first order auto-regressive setting, a similar scoring iteration is used

$$\hat{\boldsymbol{\alpha}}^{(h+1)} = \hat{\boldsymbol{\alpha}}^{(h)} + (\hat{H}^{(h)}_{\alpha\alpha'})^{-1}\frac{\partial \eta^{(h)}}{\partial \boldsymbol{\alpha}}.$$

Empirical Bayes estimates are used for the random effects $\hat{\boldsymbol{\tau}}_i$.

Goodness-of-fit was evaluated using an extension of Akaike's information criterion to robust regression, in particular to Huber's ρ function.

Using the example discussed in Section 6.1.1, these robust estimates were compared to the usual ML estimates, proving to be robust to outlying observations. In addition, the robust variance parameter estimates were lower than the corresponding ML estimates.

2.3.2 M-robust Estimation

Hu and Lachin (2001) introduced the truncated robust estimating equations, a direct extension to the GEE method, based on Huber's M-estimation.

The truncation function is given by

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \sum_{i=1}^{K} D_i^T R_i^{-1}(\boldsymbol{\alpha}) \psi(\mathbf{r}_i), \tag{2.10}$$

with $\psi(\cdot)$ the multivariate generalization to univariate M-estimation, and $R_i^{-1}(\alpha) = \text{var}(\mathbf{r}_i)$.

Note that (2.10) is different from the estimating equation (2.4) in bounding the influence of the (standardized) residuals. Hence, if ψ is the identity function, (2.10) reduces to GEE.

A further robust generalization uses $\tilde{V}_i = \text{var}(\psi(\mathbf{r}_i))$ (the variance-covariance matrix of the truncated residuals) instead of $R_i(\boldsymbol{\alpha})$ in the estimating equation (2.10). It has been shown, though, that the difference between the correlation of the truncated and untruncated residuals is trivial. Based on simulation studies, Hu chose to use $\tilde{V}_i$ when scales are heterogeneous over time.

Under the same regularity conditions imposed to GEE, with the additional assumption that $E[\psi(r)] = 0$, the truncated robust estimator of (2.10) is consistent for $\boldsymbol{\beta}$ and its asymptotic distribution is:

$$N_p(\boldsymbol{\beta}, \lim_{K \to \infty} K V_D^{-1} V_0 (V_D^{-1})^T),$$

where $V_D = \sum_{i=1}^{K} D_i^T R_i^{-1}(\boldsymbol{\alpha}) E[\psi'(\mathbf{r}_i)] D_i$, and $V_0 = \sum_{i=1}^{K} D_i^T R_i^{-1}(\boldsymbol{\alpha}) \text{cov}(\psi(\mathbf{r}_i)) R_i^{-1}(\boldsymbol{\alpha}) D_i$.

The median absolute deviation was used as an estimator of scale:

$$\hat{\phi} = b\text{median}\{|\hat{r}_{it} - \text{median}\{\hat{r}_{it}\}|\},$$

with b dependent on the distribution of the residuals, chosen such that $\hat{\phi}$ is an unbiased estimator. For example, for normally distributed data, $b = 1.483$ and for double exponential distributed data, $b = 2.1$.

Note that the estimator of $\boldsymbol{\beta}$ is consistent if $\hat{\phi}$ is unbiased, so in practice $b = 1.483$ or $b = 2.1$ is used.

A simulation was conducted for a linear model, with $b = 2.1$ and the truncated constant $c = 1.345$ to compare this estimator and GEE, under a multitude of error distributions. It was shown that the GEE is, as expected, more efficient for normal errors (relative MSE= .94) and less efficient for non-normal data, such as mixed normal error of 80% standard normal and 20% N(0,9) (relative MSE= 1.3) or mixed normal error of 95% standard normal and 5% gamma(2,4) (relative MSE= 1.69). Both estimators were unbiased and there were no differences between the two with respect to variance estimators.

The authors note that:

> One danger in the use of the robust estimating equations is to overtruncate the tail values, which may wash out some useful information in the tails. The cut-off point c in the ψ-function is often chosen such

that the robust estimating equations are 95 percent efficient when the error distribution is exactly normal.

CHAPTER III

ROBUST WEIGHTED ESTIMATION

3.1 Development for Generalized Linear Models

Assume that we have a vector of observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ following an exponential family distribution:

$$f(\mathbf{y}; \boldsymbol{\theta}, \phi) = \exp[\{\mathbf{y}^T\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\mathbf{y})\}\phi] \tag{3.1}$$

and the log-likelihood function given by

$$l(\boldsymbol{\theta}, \phi; \mathbf{y}) = \phi\{\mathbf{y}^T\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\mathbf{y})\} \tag{3.2}$$

Let $\boldsymbol{\mu} = E(\mathbf{y})$. From

$$E(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}) = E(\phi(\mathbf{y} - \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})) = E(\mathbf{y}) - \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

it follows that,

$$\boldsymbol{\mu} = \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{3.3}$$

In a similar fashion, from

$$E(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}) + E(\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}})^2 = 0,$$

we obtain the variance

$$\text{Var}(\mathbf{y}) = \sigma^2(\frac{\partial^2 a(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}) = \sigma^2 V(\boldsymbol{\mu}) \qquad (3.4)$$

Note that

$$\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}} = \frac{\partial}{\partial\boldsymbol{\theta}}(\frac{\partial a(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}) = V(\boldsymbol{\mu})$$

We are fitting the model

$$g(\mu_i) = \sum_{j=1}^{p}\beta_j x_{ij}, \quad 1 \le i \le n, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \qquad (3.5)$$

where g is a known one-to-one third order differentiable function, called the *link function.*

In vector notation, we have $g(\boldsymbol{\mu}) = X^{n\times p}\boldsymbol{\beta}$. Hence we can express the mean $\boldsymbol{\mu}$ as a function (not necessarily linear) of the parameter $\boldsymbol{\beta}$

$$\boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\beta}) = g^{-1}(X\boldsymbol{\beta}), \qquad (3.6)$$

Following the least squares theory, the maximum likelihood estimator (MLE) is the solution to:

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \phi(\mathbf{y} - \boldsymbol{\mu}) = 0,$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\mu}} &= \frac{\partial l}{\partial \boldsymbol{\theta}} \cdot \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \\ &= \phi V^{-1}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})\end{aligned}$$

Hence, the **least squares normal equations** are:

$$V^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}))(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) = 0. \tag{3.7}$$

Now let

$$D(\boldsymbol{\beta})^{n \times p} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$$

and $l(\boldsymbol{\beta}) = \mathbf{l}(\mu(\boldsymbol{\beta}), \mathbf{y})$. It is easy to show that

$$\begin{aligned}\dot{l}(\boldsymbol{\beta}) &= \phi D^T V^{-1}(\mathbf{y} - \mu(\boldsymbol{\beta})) \\ &= \phi D^T(\boldsymbol{\beta})\mathbf{V}^{-1}(\boldsymbol{\beta})(\mathbf{y} - \mu(\boldsymbol{\beta}))\end{aligned} \tag{3.8}$$

and

$$\ddot{l}(\boldsymbol{\beta}) = \phi[\mathbf{e}^{\mathbf{T}}\mathbf{V}^{-1}][\mathbf{W} - \boldsymbol{\Gamma}] - \phi \mathbf{D}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{D} \tag{3.9}$$

where $\mathbf{e} = \mathbf{y} - \mu(\boldsymbol{\beta})$, $W = \frac{\partial^2 \mu(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathbf{T}}}$, $\Gamma = D^T V^{-1} S V^{-1} D$, and $S = \frac{\partial^2 \mu(\boldsymbol{\beta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$.

Then the MLE solves:

$$\dot{l}(\hat{\boldsymbol{\beta}}) = 0$$

Using a Taylor expansion about $\boldsymbol{\beta}_0$, $\dot{l}(\hat{\boldsymbol{\beta}}) \approxeq \dot{l}(\boldsymbol{\beta}_0) + \ddot{l}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approxeq 0$. So

$$\hat{\boldsymbol{\beta}} \approxeq \boldsymbol{\beta}_0 + \{-\ddot{l}(\boldsymbol{\beta}_0)\}^{-1}\dot{l}(\boldsymbol{\beta}_0) \tag{3.10}$$

To obtain the estimator, $\hat{\boldsymbol{\beta}}$, the Gauss-Newton approximation method is being used, with the iteration step of:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \{-\ddot{l}(\boldsymbol{\beta}_0)\}^{-1}\dot{l}(\boldsymbol{\beta}^{(i)})$$

Replacing $-\ddot{l}(\boldsymbol{\beta})$ by its expectation: $E(-\ddot{l}(\boldsymbol{\beta})) = \phi D^T V^{-1} D$, Gauss-Newton becomes the iterative weighted least squares

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} + \{(D^T V^{-1} D)^{-1} D^T V^{-1} \mathbf{e}\}^{(i)} \qquad (3.11) \\ &= \{(D^T V^{-1} D)^{-1} D^T V^{-1} Z\}^{(i)} \end{aligned}$$

where $Z = D\beta + \mathbf{e}$.

This iteration is exactly the least squares solution to $Z = \mu(\boldsymbol{\beta}) + \gamma$, where $E(\gamma) = 0$ and $\text{Var}(\gamma) = V$.

Hence, the GLM model is equivalent to the regression model

$$y_i = \boldsymbol{\mu}(\boldsymbol{\beta}_0) + \epsilon_i,$$

where $Var(\epsilon_i) = V$, and $g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$.

3.2 Rank-Based Robustness and Link Functions

In general, given the model

$$\mathbf{y} = \boldsymbol{\mu}(\boldsymbol{\beta}) + \epsilon,$$

the least squares estimator minimizes

$$\|\mathbf{e}\|_{LS}^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \mu_i(\boldsymbol{\beta})]^2.$$

Equivalently, $\|\cdot\|_{LS}^2$ can be expressed as

$$\|\mathbf{e}\|_{LS}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (e_i - e_j)^2.$$

The rank-based analysis is based on the pseudo-norm

$$\|\mathbf{e}\|_R = \sum_{i=1}^{n} a_\varphi(R(e_i)) e_i,$$

where $a_\varphi(i)$ are scores such that $a_\varphi(1) \leq \cdots \leq a_\varphi(n)$ and $\sum_{i=1}^{n} a_\varphi(i) = 0$.

General rank scores have the form

$$a_\varphi(i) = \varphi(\frac{i}{n+1}),$$

where $\varphi(u)$ is a nondecreasing function defined on the interval (0,1), $\int_0^1 \varphi(u)du = 0$ and $\int_0^1 \varphi^2(u)du = 1$.

This study only considers Wilcoxon scores, generated by the function $\varphi_R(u) = \sqrt{12}(u - \frac{1}{2})$.

Hence, the rank-robust estimator minimizes

$$||\mathbf{e}||_R^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |e_i - e_j|,$$

or, equivalently, the rank-robust estimator minimizes

$$\left[\binom{n}{2}\right]^{-1} \sum_{i<j} |e_i - e_j|.$$

For the linear model, that is, when g is the *identity link*, or $\boldsymbol{\mu}(\boldsymbol{\beta}) = X\boldsymbol{\beta}$, and when e_i are iid, the asymptotic theory of the Wilcoxon estimator was developed by Jaeckel (1972). McKean and Hettmansperger (1976, 1978) developed the corresponding inference theory; see chapters 3-5 of Hettmansperger and McKean (1998) for discussion.

In practice, the form of the link is mainly dictated by previous knowledge and research with similar data. McCullagh and Nelder (1989) present a multitude of examples, mainly using discrete data, illustrating different link functions.

Besides the *identity link*, widely used in practice, the *logit link*

$$g(\boldsymbol{\mu}) = ln(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}})$$

is popular, mainly used with binomial distributed data, $Bi(n_i, \mu_i)$, $0 < \mu_i < 1$.

Other links employed with binomial data are the *probit*,

$$g(\boldsymbol{\mu}) = \Phi^{-1}(\boldsymbol{\mu})$$

and the *complementary log-log*,

$$g(\boldsymbol{\mu}) = ln[-ln(1-\boldsymbol{\mu})].$$

It has been shown that when the true mean, $\boldsymbol{\mu}$, is close to 0.5, it is hard to differentiate among these link functions.

For counts data, not in form of proportions, the *log-linear link*,

$$g(\boldsymbol{\mu}) = ln(\boldsymbol{\mu})$$

is used. The log-linear link is used with counts of events in Poisson-like processes.

For example, one might consider a model using a log-linear link if the dependent variable is the number of new cases of juvenile diabetes when collecting data over a decade.

When modeling a gamma distribution, such as the rate of a process, the *canonical link*,

$$g(\boldsymbol{\mu}) = \boldsymbol{\mu}^{-1}$$

is generally used. For example, assume the mean of a process is $\boldsymbol{\mu}(\boldsymbol{\beta}) = \frac{x}{\beta_0 x + \beta_1}$, then the *inverse linear* link is just $g(\boldsymbol{\mu}) = \beta_0 + \frac{\beta_1}{x}$.

3.3 Development for Generalized Estimating Equations

Following the same notation introduced in Section 2.1, we have $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ following an exponential family distribution:

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi].$$

We are now generalizing the GLM model to the GEE, fitting

$$g(\boldsymbol{\mu}_i) = f_i(\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \tag{3.12}$$

where g is a known link function and f_i is known.

In vector notation, we have $g(\boldsymbol{\mu}) = \mathbf{f}(\boldsymbol{\beta})$. Hence

$$\boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\beta}) = (\mathbf{g}^{-1} \circ \mathbf{f})(\boldsymbol{\beta}) = \mathbf{h}(\boldsymbol{\beta}), \tag{3.13}$$

with $\mathbf{h}(\cdot)$ nonlinear.

The development for the GLM detailed in Section 3.1 applies, using the $\boldsymbol{\mu}$ form in (3.13), instead of $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\beta})$.

It follows that the GEE model is equivalent to the nonlinear regression model

$$y_i = h_i(\boldsymbol{\beta}_0) + \epsilon_i,$$

where $Var(\epsilon_i) = V$.

3.4 Asymptotic Properties

Consider the general regression model, equivalent to the GEE,

$$y_i = h_i(\boldsymbol{\beta}_0) + \varepsilon_i, \quad \text{for} \quad 1 \leq i \leq n\,, \tag{3.14}$$

where $Var(\epsilon_i) = V$, V positive definite , where V^{-1} can be factorized as $V^{-\frac{1}{2}}V^{-\frac{1}{2}}$, and $\boldsymbol{\beta}_0 \in \boldsymbol{\Theta}^{\circ}$. We will assume that $\boldsymbol{\Theta}$ is a compact subspace of $\Re^p$.

Standardize the model in (3.14), by multiplying with $V^{-\frac{1}{2}}$, to obtain

$$y_i^* = f_i^*(\boldsymbol{\beta}_0) + \varepsilon_i^*, \tag{3.15}$$

$y_i^* = V^{-\frac{1}{2}} y_i$, $f_i^*(\boldsymbol{\beta}_0) = V^{-\frac{1}{2}} h_i(\boldsymbol{\beta}_0)$, and $\varepsilon_i^* = V^{-\frac{1}{2}} \varepsilon_i$, independent , with $Var(\varepsilon_i^*) = I^{p \times p}$.

We define the weighted Wilcoxon dispersion function by

$$D_n^w(\boldsymbol{\beta}) \equiv \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\widehat{\boldsymbol{\beta}}_n^{(0)}) |z_i(\boldsymbol{\beta}) - z_j(\boldsymbol{\beta})| \,, \tag{3.16}$$

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

Estimation based on the weighted Wilcoxon dispersion function was introduced by Sievers (1983) who assumed the weights, w_{ij}, to be non-stochastic. Naranjo and Hettmansperger (1994) further developed the weighted Wilcoxon and used it to obtain the so-called generalized R (GR) estimates of regression coefficients in the linear model. By using Mallows weights they were able to obtain estimators with a bounded influence function.

The weighted Wilcoxon dispersion function (3.16) in its present form was given by Chang et al. (1999). They considered the linear model, $f_i^*(\boldsymbol{\beta}_0) = \mathbf{x}_i^T \boldsymbol{\beta}_0$, and obtained estimates of $\boldsymbol{\beta}_0$ that have high breakdown point and at the same time possess high efficiency. They also showed that if the initial estimator, $\widehat{\boldsymbol{\beta}}_n^{(0)}$, has high breakdown, then the estimate obtained by minimizing (3.16) will have high breakdown as well; for more details, see also the discussion in Hettmansperger and McKean (1998).

In model (3.15), when the variance V of ε_i is assumed to be known, the model is a nonlinear model. For this case, we sketch the theory for the Wilcoxon estimator; see Abebe (2002) for details. We do extend the discussion in several instances, so that we can generalize the results for our later development.

The following theorem gives the existence of the minimizer of (3.16).

Theorem 3.4.1. *Under model* (3.14), *if for* $1 \le i < j \le n$, $w_{ij}(\cdot)$ *are continuous,* $\Re^+$ *valued functions, then* $\widehat{\boldsymbol{\beta}}_{V,n}$ *exists.*

In order to prove the consistency of the estimator, define $h_i(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = f_i^*(\boldsymbol{\beta}) - f_i^*(\boldsymbol{\beta}^*)$, $h_{ij}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = h_i(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - h_j(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$,
and $\Delta_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = n^{-1} \sum_{i=1}^{n} \{h_i(\boldsymbol{\beta}, \boldsymbol{\beta}^*)\}^2$.

The following assumptions will be needed.

A1: For $1 \le i, j \le n$, w_{ij} are nonnegative, continuous functions with $w_{ij}(\boldsymbol{\beta}_0) \le M < \infty$, $\sum_{i,j} w_{ij}(\boldsymbol{\beta}_0) > 0$, and gradients ∇w_{ij} bounded uniformly in i and j on $\boldsymbol{\Theta}^\circ$.

A2: $\lim_{n\to\infty} n^{-1} \Delta_n(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = 0$ for all $\boldsymbol{\beta} \in \boldsymbol{\Theta}$.

A3: $\widehat{\boldsymbol{\beta}}_n^{(0)} \to \boldsymbol{\beta}_0$ in probability.

The following lemma shows the convergence for the process D_n^w provided that the weight functions and the initial estimator, $\widehat{\boldsymbol{\beta}}_n^{(0)}$, behave in a favorable manner.

Lemma 3.4.1. *Under assumptions A1 - A3,*

$$\{D_n^w(\beta) - D_n^w(\beta_0)\} - E\{D_n^w(\beta) - D_n^w(\beta_0)\} \rightarrow 0 ,$$

in probability.

In addition to A1 - A3 above, assume the following.

A4: $\varepsilon_i^* - \varepsilon_j^*$ have a common distribution G which satisfies $G(0) = 1/2$ and has density g continuous at 0 with $g(0) > 0$.

A5: For $1 \leq i, j \leq n$, and $\boldsymbol{\Theta}^*$ a closed subset of $\boldsymbol{\Theta} \setminus \{\boldsymbol{\beta}_0\}$, there exist a $\eta > 0$ and a n_0 such that for all $n \geq n_0$ we have

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

Lemma 3.4.2. *Under A4 and A5, there exists a $\xi > 0$ and a n_0 such that for all $n \geq n_0$,*

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

The following theorem gives the consistency of $\widehat{\boldsymbol{\beta}}_{V,n}$.

Theorem 3.4.2. *Under A1 - A5, $\widehat{\boldsymbol{\beta}}_{V,n}$ is weakly consistent for $\boldsymbol{\beta}_0$.*

Proof. The proof follows from Lemma 3.4.1 and Lemma 3.4.2. □

For the asymptotic normality, a linear approximation of the model is used. Define the following errors, based on a Taylor approximation around β_0 for $f_i^*(\beta)$

$$e_i^L(\beta) = y_i^* - f_i^*(\beta_0) + \{\nabla f_i^*(\beta_0)\}^T(\beta - \beta_0) \quad for 1 \le i \le n. \tag{3.17}$$

Then the model

$$y_i^L = (\mathbf{x}_i)^T\beta_0 + \varepsilon_i^* \, , \tag{3.18}$$

has errors e_i^L, where $y_i^L = y_i^* - f_i^*(\beta_0) + \{\nabla f_i^*(\beta_0)\}^T\beta$, and $\mathbf{x}_i = \nabla f_i^*(\beta_0)$. We shall use this model several times in the sequel.

Define

$$S_n(\beta) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\hat{\beta}_n^{(0)})|e_i^L(\beta) - e_j^L(\beta)| \, . \tag{3.19}$$

Denote the minimizer of S_n by $\widehat{\widehat{\beta}}_n$.This estimator is the one considered by Chang et al. (1999) as a high breakdown estimator of linear regression coefficients.

Define the associated weighted Wilcoxon dispersion function with deterministic weights as

$$T_n(\beta) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\beta_0)|e_i^L(\beta) - e_j^L(\beta)| \, . \tag{3.20}$$

Denote the minimizer of T_n by $\widetilde{\beta}_n$. Notice that this is an estimator of a vector of linear regression coefficients. T_n corresponds to the dispersion function given by Sievers (1983). To prove the asymptotic normality of $\widehat{\beta}_{V,n}$, Abebe (2002) showed the asymptotic equivalence of $\widehat{\beta}_{V,n}$ and $\widetilde{\beta}_n$ (or, under weaker conditions, the asymptotic equivalence of $\widehat{\beta}_{V,n}$ and $\widehat{\widehat{\beta}}_n$) and then the asymptotic normality of $\widetilde{\beta}_n$.

Assume the following.

N1: The true errors, ε_i, are independent, identically distributed with

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

N2: For $1 \leq i \leq n$ and $1 \leq j \leq p$, $\nabla f^*_{ij}(\beta) = \nabla(V^{-\frac{1}{2}} h_{ij}(\beta)$ are continuous in β on Θ°.

Lemma 3.4.3. *Under A1-A5, N1, and N2*

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

in probability.

Let $\gamma = \int h_{\varepsilon^*}{}^2$ where h_{ε^*} is the density of ε^*_1. Define

$$A_{ik}(\beta_0) = \sum_{j=1}^{n} w_{ij}(\beta_0)(\nabla f^*_{ij}(\beta_0) - \nabla f^*_{jk}(\beta_0)), \quad 1 \leq k \leq p, \; 1 \leq i \leq n \, .$$

Let A_n be the $n \times p$ matrix with the (i,k)th element equal to A_{ik} and let $V_n = A_n^T A_n$. Let $F_c = (I_n - n^{-1} J_n) \nabla f^*$ be the centered $n \times p$ design matrix and let

$\overline{\nabla f^*}_k$ be the average of the kth column of ∇f^*. Here I_n is the $n \times n$ identity matrix while J_n is the $n \times n$ matrix of ones.

The following assumptions are given by Sievers (1983).

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sum_{i<j} w_{ij}(\boldsymbol{\beta}_0)(\nabla f^*_{jk}(\boldsymbol{\beta}_0) - \nabla f^*_{ik}(\boldsymbol{\beta}_0))(\nabla f^*_{jl}(\boldsymbol{\beta}_0) - \nabla f^*_{il}(\boldsymbol{\beta}_0)).$$

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

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

Theorem 3.4.3. *Under A4, SN1-SN8,*

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

Under weaker assumptions, the following lemma, equivalent to Lemma 3.4.3 can be proved (in a similar fashion):

Lemma 3.4.4. *Under A1-A5, and N2*

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

in probability.

This result will prove important for our development. It connects the theory of the estimator in the nonlinear model to the linear model (3.18).

The following theorem along with a proof can be found in Hettmansperger and McKean (1998).

Theorem 3.4.4. *Under A4, SN1-SN8,*

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\beta}}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, V_{asy}) \ .$$

The following theorems prove that the two estimators, $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_{V,n}$ are equivalent.

Theorem 3.4.5. *If*

(i) $\widehat{\boldsymbol{\beta}}_n^{(0)} \xrightarrow{\mathcal{P}} \boldsymbol{\beta}_0$ *(A3),*

(ii) $\nabla f_{ij}^*(\boldsymbol{\beta}_0)$ *are uniformly bounded for all* $1 \leq i < j \leq n$ *(N2), and*

(iii) $n^{-1} \sum \varepsilon_i^2 \to M$

then for any $\delta > 0$

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0|| \leq \delta} |S_n(\boldsymbol{\beta}) - T_n(\boldsymbol{\beta})| \xrightarrow{\mathcal{P}} 0.$$

Theorem 3.4.6. *Under A1-A5, N1, N2, SN1-SN8 and the assumptions of Theorem 3.4.5 we have*

$$n^{1/2}(\widehat{\widehat{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_n) \xrightarrow{\mathcal{P}} 0.$$

The main asymptotic result can be proved as a direct application of Slutsky's Theorem.

Theorem 3.4.7. *Under A1-A5, N1, N2, SN1-SN8,*

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

Proof. The proof is an immediate consequence of Lemma 3.4.3 and Theorem 3.4.3.

□

3.5 Estimators for the Variance-Covariance Matrix

In practice, it is assumed that V is not completely specified and, that, instead, its structure is known from prior research. Hence, a consistent estimator of V, $\hat{V}$, will be used in standardizing the model. If V is diagonal, we show in Section 3.5.1 that the resulting estimator is consistent, and follows the same asymptotic properties as the estimator $\widehat{\boldsymbol{\beta}}_{V,n}$. In Section 3.5.3, we are focusing our attention in obtaining consistent robust estimators when the variance-covariance matrix, V, is non-diagonal, but follows an AR(1) structure.

3.5.1 Diagonal Weights

In the nonlinear model (3.14),

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

$\text{Var}(\varepsilon_i) = V$, where V is known.

If V is a diagonal matrix, the model can be expressed as

$$y_i = h_i(\boldsymbol{\beta}_0) + \sigma_i \varepsilon_i^*, \quad \text{for} \quad 1 \le i \le n\,,$$

ε_i^* i.i.d, $\sigma_i > 0$. Note that, in this case, y_i is a scalar random variable. The above model corresponds to a generalized linear model with heteroscedastic errors.

Standardizing the model, as in (3.15), we obtain

$$y_i^* = f_i^*(\boldsymbol{\beta}_0) + \varepsilon_i^*,$$

where $y_i^* = \frac{1}{\sigma_i} h_i(\boldsymbol{\beta}_0)$ and $f_i^*(\boldsymbol{\beta}_0) = \frac{1}{\sigma_i} h_i(\boldsymbol{\beta}_0)$.

Let $\widehat{\boldsymbol{\beta}}_{V,n}$ denote the Wilcoxon estimator for this model.

Now consider the theory outlined in Section 3.4. Here the weight function $w_{ij}(\boldsymbol{\beta}_0) \equiv 1$. Consider the approximate linear model given in (3.17), i.e.

$$e_i^L(\boldsymbol{\beta}) = y_i^* - f_i^*(\boldsymbol{\beta}_0) + \{\nabla f_i^*(\boldsymbol{\beta}_0)\}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \quad 1 \le i \le n. \tag{3.21}$$

As in Section 3.4, denote the Wilcoxon estimator of model (3.21) by $\widehat{\widehat{\boldsymbol{\beta}}}_n$. Then, using the results in Section 3.4, we have that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \widehat{\widehat{\boldsymbol{\beta}}}_n) \xrightarrow{\mathcal{P}} 0 \tag{3.22}$$

and

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\beta}}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (\frac{1}{12\gamma^2})C^{-1}(\boldsymbol{\beta}_0)V(\boldsymbol{\beta}_0)C^{-1}(\boldsymbol{\beta}_0)). \tag{3.23}$$

Using $\sqrt{n}$ consistent estimators of σ_i, $\hat{\sigma}_i$, in standardizing the model, let

$$D^+(\boldsymbol{\beta}) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} |\hat{e}_i^L(\boldsymbol{\beta}) - \hat{e}_j^L(\boldsymbol{\beta})|; \tag{3.24}$$

and denote its minimizer by $\hat{\boldsymbol{\beta}}^+$.

Under certain conditions (see Section 3.5.2), Dixon and McKean (1996) have proven that

$$\sqrt{n}(\widehat{\widehat{\boldsymbol{\beta}}}_n - \hat{\boldsymbol{\beta}}^+) \xrightarrow{\mathcal{P}} 0, \tag{3.25}$$

Therefore, the following Lemma holds

Lemma 3.5.1.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \hat{\boldsymbol{\beta}}^+) \xrightarrow{\mathcal{P}} 0.$$

Proof. The result is a direct implication of equations (3.22) and (3.25). □

Theorem 3.5.1. *Under regularity assumptions,*

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

Proof. The result follows from the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{V,n}$ (3.23) and the asymptotic equivalence of the estimators, proven in Lemma 3.5.1. □

That is, when the errors are independent, with unknown heteroscedastic variances, and when a consistent estimator of scale is used in obtaining the robust estimator, the estimator is consistent and follows the same asymptotic distribution developed for the known variance-covariance matrix.

3.5.2 Issues in Applying Theorem 3.5.1

Our Theorem 3.5.1 serves as a basis for robust inference in the important cases where the errors are heteroscedastic but independent. These situations arise in practice quite often. The generalized linear models discussed in Section 3.1 are of this form. Other examples include conditional models based on most spherical multivariate models.

There is one aspect of the Wilcoxon fit that needs to be discussed.

In the nonlinear model, frequently there is no intercept parameter. This is always true for our weighted models. For even if there is an intercept parameter in the original model, the model that is fitted is the model after division by $\hat{\sigma}_i$, which has no intercept.

The Wilcoxon estimate, and R estimates in general, do not fit a no intercept model. That is, if the model is of the form

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad 1 \le i \le n, \tag{3.26}$$

where β does not include an intercept parameter (i.e., $\mathbf{1}$ is not in the column space of the design matrix), then R-estimates fit the model

$$y_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\boldsymbol{\beta} + e_i, \quad 1 \leq i \leq n. \tag{3.27}$$

Dixon and McKean (1996) proposed the following solutions to this problem. Instead of fitting model (3.26), fit the model

$$y_i = \alpha_0 + \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad 1 \leq i \leq n, \tag{3.28}$$

where α_0 is an intercept parameter, theoretically equal to 0. Let $\hat{\mathbf{Y}}^*$ be the R-fit of this model, where the intercept is estimated by the signed-rank location estimate based on the residuals $y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^*$. Let $X = [\mathbf{x}_i']$ denote the design matrix for model (3.26). Let

$$\hat{\mathbf{Y}} = P_X \hat{\mathbf{Y}}^*,$$

where for any matrix A, P_A is the projection operator onto the space spanned by the column of A, and let $\hat{\boldsymbol{\beta}}$ be the solution to

$$X\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}}.$$

Then under the additional assumption of symmetry of the error distribution, Theorem 3.5.1 holds.

If the assumption of symmetry is unrealistic, then we can proceed by fitting the intercept in model (3.28) by the median of the residuals. This affects the asymptotic variance of our estimate. By algebraic manipulation, we can show that the asymptotic variance is

$$((X^*)^T X^*)^{-1}(X^*)^T[(\frac{1}{2h_\varepsilon(0)})^2 P_1 + \tau^2 P_{X_C^*}]X^*((X^*)^T X^*)^{-1}, \tag{3.29}$$

where, for the asymptotic variance of $\widehat{\widehat{\beta}}_n$, $X^* = \nabla f^*(\beta_0)$ and $X_C^* = X^* - \bar{X}^*$. In the representation (3.29), P represents the projection matrix.

3.5.3 Non-Diagonal Weights

We will focus our attention on obtaining robust estimators, $\hat{V}$, of the variance-covariance matrix V, when V has an AR(1) structure.

In many clinical trials, subjects are followed over a period of time and the observations on each individual patient compose a time series. For these observations assumed to follow an AR(1) structure, $V = \{\sigma_{ij}\}$, where

$$\sigma_{ii} = \sigma^2, \quad \text{and} \quad \sigma_{ij} = \sigma^2 \rho^{|i-j|}, \tag{3.30}$$

where ρ is the *autocorrelation coefficient.*

The present research uses an adaptation of the robust estimate for the autocorrelation proposed by Koul and Saleh (1993), as follows.

Assuming that within each subject i, $1 \leq i \leq n$, observations follow an autoregressive series of order 1,

$$y_{it} = \rho y_{i(t-1)} + \epsilon_{it}, \quad 2 \leq t \leq p,$$

the autoregressive model can be described as a linear model

$$\mathbf{y}_i = \mathbf{x}_i \rho_i + \epsilon_i,$$

where $\mathbf{y}_i = (y_{i2}, \ldots, y_{ip})^T$, $\mathbf{x}_i = (y_{i1}, \ldots, y_{i(p-1)})^T$, and ϵ_i iid.

Hence, the estimator ρ_i is the argument which minimizes $\sum_{k<l} |e_{ik} - e_{il}|$, where $e_{ik} = (\mathbf{y}_i)_k - (\mathbf{x}_i \rho_i^{(0)})_k$, and $\rho_i^{(0)}$ an initial estimator.

Since the variance-covariance matrix V is common for all subjects, the above algorithm is applied for each subject and an overall robust estimate of ρ is obtained,

$$\hat{\rho} = \text{median}\{\hat{\rho}_1, \ldots, \hat{\rho}_n\},$$

to be used in standardizing the model (3.14).

Further research is needed in proving a result similar to Theorem 3.5.1 for non-diagonal weights, i.e. that the estimators are consistent and have the same asymptotic distribution when $\hat{\rho}$ is used as an estimator of ρ.

Terpstra et al. (2001) proposed a weighted rank-based (GR) estimate of ρ, which might provide more protection against points of high leverage. Furthermore, Terpstra et al. (2000) introduced the concept of HBR-estimates, using Schweppe-type weights, to autoregressive models.

If the time between measurements is not equally spaced, Diggle (1988) proposed an extension to the AR(1) structure of the form

$$\sigma_{ij} = \sigma^2 e^{\alpha h},$$

where h is the time distance between the two measurements, $|t_i - t_j|$.

When the covariance between two observations on the same subjects does not depend on the times of measurement but rather on the conditions under which the measurements were taken, the variance-covariance structure is assumed to be that of *compound symmetry*. In this case, observations within each patient have a variance-covariance matrix of the form

$$V = \sigma^2 J_p + \sigma_e^2 I_p,$$

where J_p is a $p \times p$ matrix of 1. That is, measurements within the same subject have the same covariance, regardless of how far apart (in time) they were collected. More research is needed in obtaining robust estimators of the compound symmetry matrix.

CHAPTER IV

ASYMPTOTIC TESTS

Although most of the theory developed for the Generalized Estimation has focused on estimation, in practice there is often the case that testing is needed. Often, more than one treatment group is being followed over time, for example, and differences among treatments (or some other type of contrasts) are subject to investigation.

We are developing the testing theory for the Generalized Estimating Equations, using the robust theory presented in Chapter III.

4.1 Wald-Type and Score Tests

In order to test the hypotheses

$$H_O : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{vs} \quad H_a : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \tag{4.1}$$

we developed a robust equivalent to Wald's test.

Since

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, (\frac{1}{12\gamma^2})C^{-1}(\boldsymbol{\beta}_0)V(\boldsymbol{\beta}_0)C^{-1}(\boldsymbol{\beta}_0))$$

it follows that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T[(\frac{1}{12\gamma^2})C^{-1}(\boldsymbol{\beta}_0)V(\boldsymbol{\beta}_0)C^{-1}(\boldsymbol{\beta}_0)]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

asymptotically has a chi-square distribution with p degrees of freedom.

Let τ_φ be the scale parameter, given by

$$\tau_\varphi^{-1} = \int \varphi(u)\varphi_{h_{\varepsilon^*}}(u),$$

$$\varphi_{h_{\varepsilon^*}}(u) = -\frac{h'_{\varepsilon^*}(H_{\varepsilon^*}^{-1}(u))}{h_{\varepsilon^*}(H_{\varepsilon^*}^{-1}(u))},$$

where h_{ε^*} is the distribution of the errors, and for our case, the Wilcoxon score function, $\varphi(u) = \sqrt{12}(u - \frac{1}{2})$, is used.

Koul, Sievers and McKean (1987) proposed a density-type estimator of τ_φ, based on residuals; see details and discussion in Section 3.7 of Hettmansperger and McKean (1998). It has been shown that, under regularity conditions, $\hat{\tau}_\varphi$ is a consistent estimator of scale.

Using calculus manipulations, it is easy to show that

$$\tau = \frac{1}{\sqrt{12}\gamma}.$$

Then the Wald's test F_W , given by

$$F_W = \frac{[(\hat{\beta}-\beta_0)^T(C^{-1}(\beta_0)V(\beta_0)C^{-1}(\beta_0))^{-1}(\hat{\beta}-\beta_0)]/p}{\hat{\tau}^2}, \qquad (4.2)$$

follows a chi-square distribution with p degrees of freedom. Small sample studies suggest using the F distribution with p and $n-p-1$ degrees of freedom as an approximation to the asymptotic chi-square distribution.

Similarly, to test the general linear model hypothesis

$$H_O : M\beta = 0 \quad \text{vs} \quad H_a : M\beta \neq 0,$$

where $M^{q\times p}$ matrix of constraints, one would use the Wald's test statistic given by

$$F_W = \frac{[(M\hat{\beta})^T(MC^{-1}(\beta_0)V(\beta_0)C^{-1}(\beta_0)M^T)^{-1}(M\hat{\beta})]/q}{\hat{\tau}^2}.$$

The approximate test would be to compute this test with $F_W \sim F(q, n-p-1)$ critical values.

Another robust test is the rank gradient scores test. This is asymptotically equivalent to the Wald's test described in (4.2) for testing the hypothesis (4.1).

Given the distribution of the gradient in the linear model setting (see Hettmansperger and McKean (1998), page 174, for details)

$$S(\beta_0) \xrightarrow{\mathcal{D}} N_p(0, n\Sigma),$$

the test

$$F_R = \frac{1}{n} S(\boldsymbol{\beta}_0)^T \Sigma^{-1} S(\boldsymbol{\beta}_0) \tag{4.3}$$

follows a chi-square distribution with p degrees of freedom (unlike the Wald type test, there is no natural F-approximation).

If we replace Σ by $\frac{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))}{n}$ and $S(\boldsymbol{\beta}_0)$ by $\sqrt{12}(\nabla f^*(\boldsymbol{\beta}_0))^T[H(\mathbf{e}^*) - \frac{1}{2}]$, (4.3) simplifies to

$$F_R = 12(H(\mathbf{e}^*) - \frac{1}{2})^T P (H(\mathbf{e}^*) - \frac{1}{2}),$$

with P, projection onto the tangent plane, defined as

$$(\nabla f^*(\boldsymbol{\beta}_0))\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-1}(\nabla f^*(\boldsymbol{\beta}_0))^T.$$

4.2 Drop in Dispersion Test

Recall the linear model in (3.18)

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

Since this is a linear model with i.i.d errors, the linear models theory described in Chapter 3 of Hettmansperger and McKean (1998) directly applies. Hence, the drop in dispersion test for testing the hypotheses introduced in (4.1)

$$H_O : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{vs} \quad H_a : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$$

is given by

$$F_S = \frac{RD/p}{\hat{\tau}/2}, \tag{4.4}$$

where $RD = S_n(\boldsymbol{\beta}_0) - S_n(\boldsymbol{\beta})$, and S_n given in (3.19),

$$S_n(\boldsymbol{\beta}) = \left[\binom{n}{2}\right]^{-1} \sum_{i<j} w_{ij}(\hat{\boldsymbol{\beta}}_n^{(0)})|e_i^L(\boldsymbol{\beta}) - e_j^L(\boldsymbol{\beta})|,$$

and $w_{ij}(\hat{\boldsymbol{\beta}}_n^{(0)}) \equiv 1$. The drop in dispersion test follows an asymptotic χ^2 distribution, but Hettmansperger and McKean (1998) have shown in small-sample studies that it is best to compare F_S to F critical values with p and $n-p-1$ degrees of freedom.

The test can be used with the non-weighted dispersion function, and the examples in Chapter VI illustrate its applicability when the errors are correlated and an estimator of the variance-covariance matrix V is used in standardizing the model.

4.3 Testing for Different Link Functions

Although, often, the link function is known from historical research, it is interesting to develop a test of accuracy of the link function. In many instances, the general family of the link function is known, but choosing among members of that family is a task many times left on the hands of the clinician, not the statistician.

Using the previous notation, testing for a link function can be written as

$$H_O : h(\boldsymbol{\beta}) = h_0(\boldsymbol{\beta}) \quad \text{vs} \quad H_a : h(\boldsymbol{\beta}) \neq h_0(\boldsymbol{\beta}), \tag{4.5}$$

where $h(\boldsymbol{\beta})$ and $h_0(\boldsymbol{\beta})$ are different link functions.

Notice that (4.5) can be written as

$$H_O : d(\boldsymbol{\beta}) = 0 \quad \text{vs} \quad H_a : d(\boldsymbol{\beta}) \neq 0,$$

where $d(\boldsymbol{\beta}) = h(\boldsymbol{\beta}) - h_0(\boldsymbol{\beta})$, $d : \Re^p \to \Re^q$.

Under this setup, (4.5) will be tested using a Wald-type test derived in a similar fashion as (4.2), i.e.,

$$F_W = \frac{[d(\hat{\boldsymbol{\beta}})^T(\hat{H}_d C^{-1}(\boldsymbol{\beta}_0)V(\boldsymbol{\beta}_0)C^{-1}(\boldsymbol{\beta}_0)\hat{H}_d^T)^{-1}d(\hat{\boldsymbol{\beta}})]/q}{\hat{\tau}^2}, \tag{4.6}$$

where H_d is the jacobian matrix of d.

Thus an approximate size α test for $d(\boldsymbol{\beta}) = 0$ is

$$\phi(F) = \begin{cases} 1 & \text{if } F > F^{\alpha}_{q,n-p-1} \\ 0 & \text{if } F \leq F^{\alpha}_{q,n-p-1} \end{cases}.$$

4.4 Discussion

There has been extensive discussions in the literature regarding GLM and the interpretation of the tests for the standardized model. That is, if the model is of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon.$$

$Var(\varepsilon) = V$, should we regress $\mathbf{Y}^*$ on $\mathbf{X}^*$, or should we use the original design matrix and regress $\mathbf{Y}^*$ on $\mathbf{X}$? Here, $\mathbf{Y}^*$ and $\mathbf{X}^*$ correspond to the standardized model, $\mathbf{Y}^* = V^{-\frac{1}{2}}\mathbf{Y}$, and $\mathbf{X}^* = V^{-\frac{1}{2}}\mathbf{X}$.

In simulation studies using the multiple sclerosis data, we found the two methods to be similar (see Table 2), and we have chosen to present the results for the fully standardized model (regressing $\mathbf{Y}^*$ on $\mathbf{X}^*$) for the remainder of this study. Details of the simulation study are presented in Section 6.1.2.

Table 2

Wilcoxon Tests: Standardized vs. Non-standardized Design Matrix

Wilcoxon: Test	F test, using $\mathbf{X}^*$	F test, using $\mathbf{X}$
Equal intercepts and slopes	9.0328	9.6237
Equal intercepts	1.9784	1.7428
Equal slopes	1.5982	1.4456

CHAPTER V

DIAGNOSTICS

An important part of the analysis is the examination of the fitted values. Besides residual plots, overall and at each time point, another useful tool are diagnostic techniques. We are focusing on diagnostic techniques that detect outlying cases and influential subjects.

5.1 Asymptotic Representation of $\widehat{\boldsymbol{\beta}}_{V,n}$

As shown in Theorem 3.4.7, the asymptotic distribution of the estimator is given by

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

Since $S(\boldsymbol{\beta}_0)\xrightarrow{\mathcal{D}}N_p(0, n\Sigma)$, let the following be the asymptotic representation of $\widehat{\boldsymbol{\beta}}_{V,n}$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{12}\gamma}\{C^{-1}(\boldsymbol{\beta}_0)V(\boldsymbol{\beta}_0)C^{-1}(\boldsymbol{\beta}_0)\}^{\frac{1}{2}}\Sigma^{\frac{-1}{2}}\frac{S(\boldsymbol{\beta}_0)}{\sqrt{n}}, \qquad (5.1)$$

where $\Sigma = \lim_{n\to\infty} \frac{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))}{n}$,

and $S(\boldsymbol{\beta}_0) = \sqrt{12}(\nabla f^*(\boldsymbol{\beta}_0))^T\{H[\mathbf{y}^L - (\nabla f^*(\boldsymbol{\beta}_0))\boldsymbol{\beta}_0] - \frac{1}{2}\}$.

Using SN1 and SN8, (5.1) simplifies further to

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{12}\gamma}\{n^4 C_n^{-1}(\boldsymbol{\beta}_0) n^{-3} V_n(\boldsymbol{\beta}_0) C_n^{-1}(\boldsymbol{\beta}_0)\}^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \frac{S(\boldsymbol{\beta}_0)}{\sqrt{n}} + o_p(1)$$

Denote $B = nC_n^{-1}(\boldsymbol{\beta}_0)V_n(\boldsymbol{\beta}_0)C_n^{-1}(\boldsymbol{\beta}_0)$. Then

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{V,n} - \boldsymbol{\beta}_0) &= \frac{\gamma^{-1}}{\sqrt{n}}(B)^{\frac{1}{2}}\left\{\frac{(\nabla f^*(\beta_0))^T(\nabla f^*(\beta_0))}{n}\right\}^{-\frac{1}{2}} \times \\ &\quad \times (\nabla f^*(\boldsymbol{\beta}_0))^T\{H[\mathbf{y}^L - \nabla f^*(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0] - \frac{1}{2}\} + o_p(1) \\ &= \gamma^{-1}(B)^{\frac{1}{2}}\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-\frac{1}{2}} \times \\ &\quad \times (\nabla f^*(\boldsymbol{\beta}_0))^T\{H[\mathbf{y}^L - \nabla f^*(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0] - \frac{1}{2}\} + o_p(1)\end{aligned}$$

Note that in a neighborhood of $\boldsymbol{\beta}_0$, $H[\mathbf{y}^L - \nabla f^*(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0]$ is approximately equal to $H(\mathbf{e}^*)$.

Hence,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{V,n} &= \boldsymbol{\beta}_0 + \frac{\gamma^{-1}}{\sqrt{n}}(B)^{\frac{1}{2}}\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-\frac{1}{2}} \times \\ &\quad \times (\nabla f^*(\boldsymbol{\beta}_0))^T[H(\mathbf{e}^*) - \frac{1}{2}] + o_p(n^{-\frac{1}{2}}) \qquad (5.2)\end{aligned}$$

In order to obtain the asymptotic representation of $\hat{\mathbf{y}}^*$, a Taylor expansion for $f^*(\boldsymbol{\beta})$, about $\boldsymbol{\beta}_0$, will be used.

$$f^*(\boldsymbol{\beta}) = f^*(\boldsymbol{\beta}_0) + (\nabla f^*(\boldsymbol{\beta}_0))^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + R_n,$$

where $R_n \to 0$ as $n \to \infty$.

Then the predicted values , $\hat{\mathbf{y}}^*$, can be expressed as

$$\begin{aligned}\hat{\mathbf{y}}^* &= f^*(\widehat{\boldsymbol{\beta}}_{V,n}) \\ &= f^*(\boldsymbol{\beta}_0) + (\nabla f^*(\boldsymbol{\beta}_0))^T(\widehat{\boldsymbol{\beta}}_{V,n} - \boldsymbol{\beta}_0) + R_n,\end{aligned}$$

where $R_n \to 0$ as $n \to \infty$.

Define

$$P^a = \frac{\gamma^{-1}}{\sqrt{n}}(\nabla f^*(\boldsymbol{\beta}_0))(B)^{\frac{1}{2}}\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-\frac{1}{2}}(\nabla f^*(\boldsymbol{\beta}_0))^T$$

and

$$P = (\nabla f^*(\boldsymbol{\beta}_0))\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-1}(\nabla f^*(\boldsymbol{\beta}_0))^T$$

Using 5.2, it follows that

$$\hat{\mathbf{y}}^* = f^*(\boldsymbol{\beta}_0) + P^a[H(\mathbf{e}^*) - \frac{1}{2}] + o_p(n^{-\frac{1}{2}}) \tag{5.3}$$

Since $\hat{\mathbf{e}}^* = \mathbf{y}^* - \hat{\mathbf{y}}^*$, it follows that the asymptotic representation of the residual is

$$\begin{aligned}\hat{\mathbf{e}}^* &= (f^*(\boldsymbol{\beta}_0) + \mathbf{e}^*) - (f^*(\boldsymbol{\beta}_0) + P^a[H(\mathbf{e}^*) - \frac{1}{2}]) \\ &= \mathbf{e}^* - P^a[H(\mathbf{e}^*) - \frac{1}{2}] + o_p(n^{-\frac{1}{2}}) \qquad (5.4)\end{aligned}$$

Suppose the true model is

$$\mathbf{y}^* = f^*(\boldsymbol{\beta}) + g(\lambda) + \epsilon^*$$

but instead, the following model is being fit

$$\begin{aligned}\mathbf{y}^* &= f^*(\boldsymbol{\beta}) + \epsilon^* \\ &= f^*(\boldsymbol{\beta}_0) + (\nabla f^*(\boldsymbol{\beta}_0))^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \epsilon^*. \end{aligned} \tag{5.5}$$

Under this misspecified model, the asymptotic representation of $\widehat{\boldsymbol{\beta}}_{V,n}$ is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{V,n} &= \boldsymbol{\beta}_0 + \frac{\gamma^{-1}}{\sqrt{n}}(B)^{\frac{1}{2}}\{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-\frac{1}{2}}(\nabla f^*(\boldsymbol{\beta}_0))^T[H(\mathbf{e}^*) - \frac{1}{2}] \\ &+ \{(\nabla f^*(\boldsymbol{\beta}_0))^T(\nabla f^*(\boldsymbol{\beta}_0))\}^{-1}(\nabla f^*(\boldsymbol{\beta}_0))^T g(\lambda) + o_p(n^{-\frac{1}{2}}) \end{aligned} \tag{5.6}$$

Note that the asymptotic representation of the Wilcoxon estimator differs from that of the LS estimator only in using a bounded function of the residuals, $H(\mathbf{e}^*) - \frac{1}{2}$.

5.2 Standardized Residuals

From the asymptotic representation of the estimator we obtain the first-order expressions of the residuals and fitted values.

$$\hat{\mathbf{y}}^* = f^*(\boldsymbol{\beta}_0) + P^a[H(\mathbf{e}^*) - \frac{1}{2}] + Pg(\lambda) + o_p(n^{-\frac{1}{2}}) \tag{5.7}$$

$$\begin{aligned}\hat{\mathbf{e}}^* &= (f^*(\boldsymbol{\beta}_0) + g(\lambda) + \mathbf{e}^*) - (f^*(\boldsymbol{\beta}_0) + P^a[H(\mathbf{e}^*) - \frac{1}{2}] + Pg(\lambda)) \\ &= e^* - P^a[H(\mathbf{e}^*) - \frac{1}{2}] + (I - P)g(\lambda) + o_p(n^{-\frac{1}{2}}) \end{aligned} \tag{5.8}$$

Note that the residuals have the expected asymptotic bias, $(I - P)g(\lambda)$.

In order to obtain the standardized residuals, the variance-covariance matrix of the residual in (5.4) will be computed. Note that

$$\begin{aligned}\mathrm{E}(\hat{\mathbf{e}}^*) &= \mathrm{E}(\mathbf{e}^* - P^a[H(\mathbf{e}^*) - \frac{1}{2}]) \\ &= \mathrm{E}(\mathbf{e}^*)\end{aligned} \tag{5.9}$$

It follows that

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{e}}^*) &= \mathrm{E}[(\hat{\mathbf{e}}^* - \mathrm{E}(\mathbf{e}^*))[(\hat{\mathbf{e}}^* - \mathrm{E}(\mathbf{e}^*))^T] \qquad (5.10)\\ &= \mathrm{E}[((\mathbf{e}^* - \mathrm{E}(\mathbf{e}^*)) - P^a(H(\mathbf{e}) - \frac{1}{2}))((\mathbf{e}^* - \mathrm{E}(\mathbf{e}^*)) - P^a(H(\mathbf{e}^*) - \frac{1}{2}))^T]\\ &= \mathrm{Cov}(\mathbf{e}^*) - 2P^a\mathrm{E}[((\mathbf{e}^* - \mathrm{E}(\mathbf{e}^*))(H(\mathbf{e}^*) - \frac{1}{2})^T] + P^a\mathrm{Cov}(F(\mathbf{e}^*))(P^a)^T\\ &= \Sigma - 2\delta P^a - \frac{1}{12}P^a(P^a)^T,\end{aligned}$$

where $\delta = \mathrm{E}[\mathbf{e}^*(H(\mathbf{e}^*) - \frac{1}{2})^T]$.

For diagnostics,we will use the standardized residuals, defined as

$$r_{it} = \frac{\hat{e}^*_{it}}{\mathrm{Var}^{\frac{1}{2}}(\hat{e}^*_{it})} \qquad (5.11)$$

We will declare a case a potential outlier if the absolute value of the standardized residual is greater than 2.

$$y_{it} \quad \text{potential outlier if} \quad |r_{it}| > 2.$$

5.3 Influential Subjects

Other statistics used with Least Squares methods in determining influential observations are Cook's (1977) measure (the change in fitted values when observations from one individual, y_i, are removed) and the DBETAS proposed by Belsley et al. (1980) (the change in the estimator values when observations from one individual, y_i, are removed).

We are extending Belsley's measure to the robust estimator and look at influential subjects, as opposed to influential cases. The method is necessary when one identifies from profile plots that the observations from one subject are substantially different from the other subjects in the study. In that case, a qq-plot (or residual plot) of overall residuals is not helpful, since the repeated observations for that subject might all lie in the tail of the distribution, suggesting an overall poor fit. Instead, a better plotting approach is creating residual plots stratified by time points.

Since influential points don't necessarily have large residuals and points with large residuals are not necessarily influential, Belsley's statistics looks at the change in the estimator values when observations from one individual, y_i, are removed, as a measure of the effect of y_i.

$$WDBETA_i = \frac{(\widehat{\beta}^{(i)}_{V,n} - \widehat{\beta}_{V,n})^T (C^{-1}(\widehat{\beta}_{V,n}) V(\widehat{\beta}_{V,n}) C^{-1}(\widehat{\beta}_{V,n}))^{-1} (\widehat{\beta}^{(i)}_{V,n} - \widehat{\beta}_{V,n})}{\hat{\tau}^2},$$

with C, V, and $\hat{\tau}$ introduced in Chapter IV, and $\widehat{\beta}^{(i)}_{V,n}$ the robust Wilcoxon estimator when observations for subject y_i are removed from the model.

In order to declare a subject to be influential, $WDBETA_i$ will be compared to $F(p, n-p-1)$.

An analogous method of identifying influential subjects is the jack-knife or cross-validatory residual, which fits the model excluding subject y_i, and standardizes the residuals using the fits and standard deviations based on this reduced model. However, it is easy to show that the jack-knife is a monotone function of the standardized residuals, and it will not be pursued any further in this study.

CHAPTER VI

EXAMPLES AND SIMULATION RESULTS

6.1 Multiple Sclerosis Example

6.1.1 Estimation and Testing

Going back to the motivating example introduced in Section 1.2, the measure of change within patient is the *log relative burden*, the logarithm of the ratio of the area of the brain impacted by Multiple Sclerosis at a given time to the area impacted by the disease at baseline. For modeling, the variance-covariance matrix of within patient observations are assumed to follow an AR(1) structure. In addition, no Wilcoxon weights will be used (i.e., $w_{ij} \equiv 1$). Only 44 patients with complete records will be used for our analysis.

The autocorrelation coefficient is estimated using the robust procedure to be 0.28037, coming close to the LS estimator of .23472, with individual AR coefficients ranging from -.347 to .916.

Outliers are present in the placebo group, at time point 6, and in the low dose group, at time 15 (see Figure 2). If the outlying points are ignored, the slopes and the intercepts for the three treatment groups seem to be equal.

The q-q plots for data (Figure 3), by treatment group, flag the above

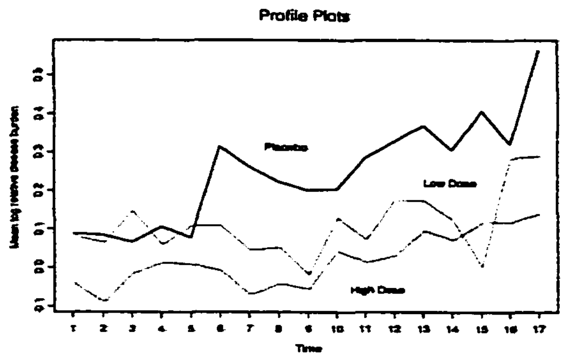

Figure 2. Profile Plots of Log-Transformed Multiple Sclerosis Data

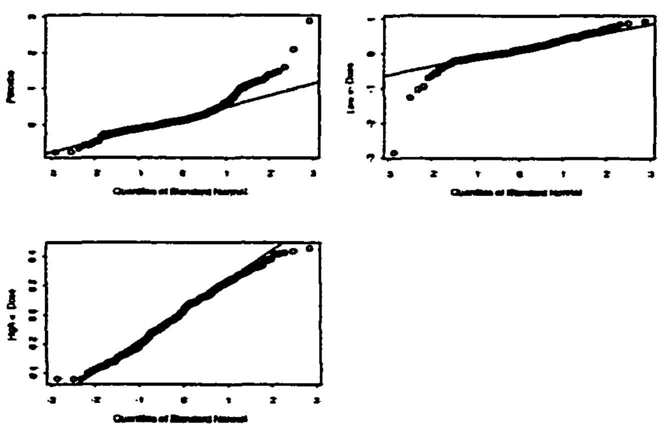

Figure 3. Q-Q Plots of Log-Transformed Multiple Sclerosis Data, by Group

mentioned observations in the placebo and low-dose group as outliers, while the high-dose group follows a short-tailed distribution.

Fitting a model of different intercept and different slopes for the three treatment groups, the Wilcoxon and Least Squares estimators introduced in Chapters II and III differ, as expected, for the skewed groups (placebo and low dose) (Table 3).

Table 3

Wilcoxon Estimates Versus LS Estimates (Multiple Sclerosis Data)

Group	Estimator	Wilcoxon	Least Squares
Placebo	Intercept	-.00542 (.03372)	.02956 (.04884)
	Slope	.01957 (.00326)	.02449 (.00473)
Low dose	Intercept	.00817 (.04936)	.00851 (.0715)
	Slope	-.00631 (.00477)	-.0161 (.00692)
High Dose	Intercept	-.0788 (.04936)	-.11268 (.0715)
	Slope	-.00745 (.00477)	-.01303 (.00692)
Overall	$\hat{\rho}$	.28037	.23472
	Scale	$\hat{\tau} = 1.80657$	$\hat{\sigma} = 1.48336$

Testing different models, for equal intercepts and slopes, equal intercepts and then equal slopes, as expected, Least Squares comes closer than Wilcoxon to declaring the three treatment slopes to be significantly different (Table 4).

Table 4

Wilcoxon Tests Versus LS Tests (Multiple Sclerosis Data)

Wilcoxon: Test	DF	F test	p-value
Equal intercepts and slopes	(4,741)	9.765	$< .001$
Equal intercepts	(2,741)	1.9784	.278
Equal slopes	(2,741)	1.5982	.406
LS : Test	DF	F test	p-value
Equal intercepts and slopes	(4,742)	12.4976	$< .001$
Equal intercepts	(2,742)	1.7084	.3638
Equal slopes	(2,742)	3.1127	.0902

Q-q plots of the standardized robust residuals (Figure 4) are comparable to

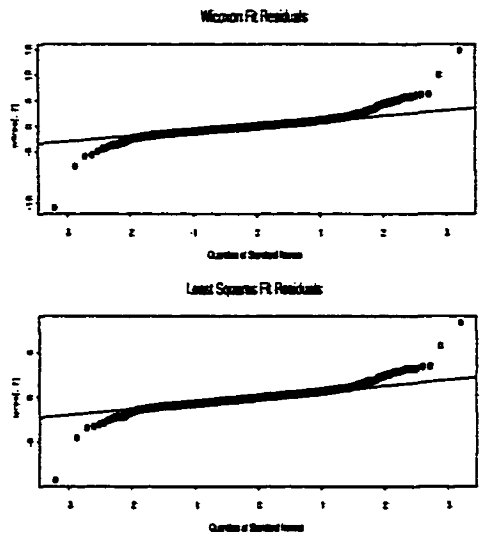

Figure 4. Residual Plots for Multiple Sclerosis Data

the standardized least squares residuals. The robust estimators are possibly biased due the biasness introduced by the AR(1) parameter estimators (see discussion in Section 6.1.2).

6.1.2 Simulation Study

We conducted a small simulation studies to see how liberal the inference is for the GEE model with autoregressive correlation structure.

Consider a linear model with time series errors. (In our case, this would be the model for one patient).

The following is often called the *Prais-Winsten* Method.

1. Fit an initial short autoregressive time series.

2. Estimate autoregressive parameter (for LS):

$$\widehat{\rho}_{LS} = \frac{\sum r_t r_{t-1}}{\sum r_t^2}\,.$$

3. Transform using $\widehat{\rho}_{LS}$ as our estimate of ρ.

4. Refit.

The resulting inference based on the second fit (item 4) is quite liberal; see, studies by McKnight et al. (1999) and Huitema et al. (1999).

For example: empirical levels for 4 effects on one such study ($n = 30$, $p = 4$, $AR(1)$, nominal $\alpha = .05$, 5000 simulations) exceed the .05 nominal level even when the true parameter, ρ, is 0. In addition, the departure from the empirical level grows with ρ.

The simulation study (500 simulations) based on the multiple sclerosis data has a total sample size of n =748, with p =6. We are assuming that all linear model parameters are 0, that the errors are normal and that $\rho \in \{0, .95\}$.

The empirical levels for pairwise slope and intercept differences are presented in Table 5 and Table 6. Wilcoxon, though liberal, is closer to the nominal levels than the LS estimators.

Table 5

Empirical α Levels of Slope Effects

	Slope Effects, $\rho = 0$			
	Slope(2) − Slope(1)		Slope(3) − Slope(1)	
α	LS	Wil.	LS	Wil.
0.10	.122	.108	.126	.114
0.05	.064	.058	.068	.060
	Slope Effects, $\rho = .95$			
	Slope(2) − Slope(1)		Slope(3) − Slope(1)	
α	LS	Wil.	LS	Wil.
0.10	.260	.240	.246	.218
0.05	.182	.144	.180	.152

Table 6

Empirical α Levels of Intercept Effects

	Intercept (Level) Effects, $\rho = 0$			
	$\mu(2) - \mu(1)$		$\mu(3) - \mu(1)$	
α	LS	Wil.	LS	Wil.
0.10	.120	.104	.120	.134
0.05	.074	.054	.068	.080
	Intercept (Level) Effects, $\rho = .95$			
	$\mu(2) - \mu(1)$		$\mu(3) - \mu(1)$	
α	LS	Wil.	LS	Wil.
0.10	.436	.404	.394	.398
0.05	.346	.328	.326	.296

The empirical levels for the same tests described in Section 6.1.1 are presented in Table 7 and Table 8.

Table 7

Empirical Levels: $\rho = 0$

	Same Model		Same Intercepts		Same Slopes	
α	LS	Wil.	LS	Wil.	LS	Wil.
0.10	.136	.130	.134	.124	.128	.128
0.05	.088	.078	.070	.064	.058	.062

Table 8

Empirical Levels: $\rho = .95$

	Same Model		Same Intercepts		Same Slopes	
α	LS	Wil.	LS	Wil.	LS	Wil.
0.10	.792	.734	.506	.472	.320	.294
0.05	.718	.660	.404	.378	.210	.200

These analyses are somewhat liberal. This is not surprising due to the liberalness of the *Prais-Winsten* procedure on linear models with time series errors. As expected, the liberalness seems to grow worse as ρ increases.

6.2 Cholesterol Example

The next example consists of cholesterol levels over 10 years for 22 randomly selected individuals from the Framingham study, analyzed by Zhang and Davidian (2001). Data is collected at baseline and then at years 2, 4, 6, 8, and 10 of the study, for a total of 6 time points. Age is a possible covariate for this study, and besides changes in cholesterol levels over time, it is of interest to compare cholesterol levels and their rate of change between genders.

As with the multiple sclerosis example, it is assumed that the longitudinal data is correlated, and that the variance-covariance matrix has an AR(1) structure. The robust estimate of the autocorrelation coefficient is .2336, while the LS estimate is .0852. The difference lies in that the robust estimate is using the median of AR coefficients obtained from individual subjects, while the LS is using the mean of the same estimators. When investigated, the distribution of the 22 AR estimators appears to be negatively skewed, with a range from -.9310 to .7222.

Plots of the data by time point (Figure 5) and the qqplots by time point (Figure 6) indicate that the data is positively skewed, with subject number 19 having significantly higher measurements at all time points.

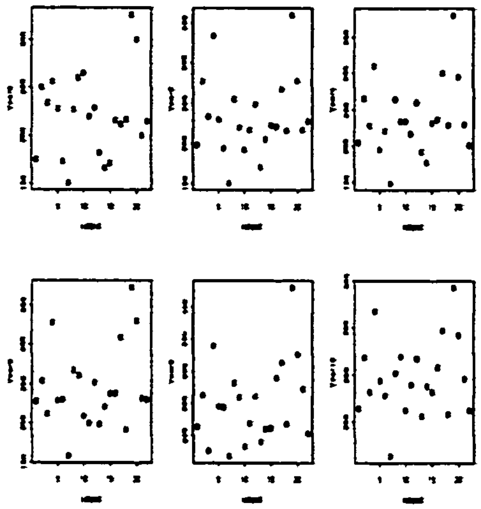

Figure 5. Scatter Plots for Cholesterol Data, by Time

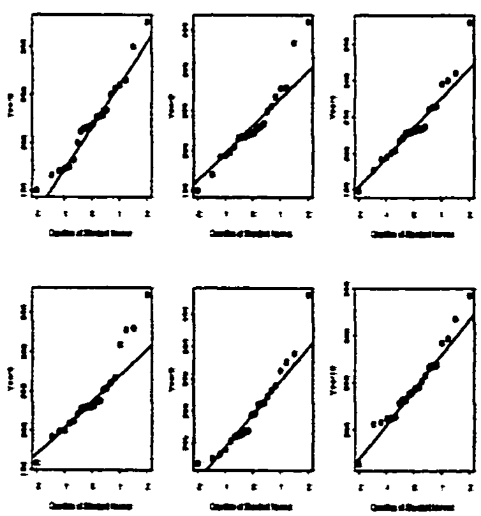

Figure 6. Q-Q Plots for Cholesterol Data, by Time

The Wilcoxon estimators are close to their Least Square counterparts (see Table 9), but their standard errors are consistently smaller. For this model, the Wilcoxon test flags a significant difference between genders (F test is 18.8739, with a p-value $< .0001$, when compared to F critical values of 1 and 128 degrees of freedom), and similar result for the LS testing (F test $= 18.0508$, $p < .0001$).

To test the robustness of the estimator, the baseline observation of the first subject was changed by a magnitude of 1. The Wilcoxon estimators were robust to this change, while the LS estimators were greatly influenced by it (see Table 10). In addition, while the Wilcoxon test of gender differences remained significant ($F = 17.6357$, $p < .0001$), the LS p-value changed from $< .0001$ to .6883 ($F = 0.1671$).

Table 9

Wilcoxon Estimates Versus LS Estimates (Cholesterol Data)

Estimator	Wilcoxon	Least Squares
Intercept	119.301 (8.29945)	118.849 (26.1041)
Gender Slope	-35.8443 (2.44994)	-32.7388 (7.70576)
Age Slope	3.3273 (.19715)	3.395 (.62009)
$\hat{\rho}$	.2336	.0852
Scale	$\hat{\tau} = 5.43807$	$\hat{\sigma} = 3.00122$

We noted in the beginning of the example that subject 19 seems to consistently have higher measurement than the other subjects. In order to analyze whether subject 19 is influential, the model was refit with 21 subjects (eliminating

Table 10

Wilcoxon Estimates Versus LS Estimates (Stressed Cholesterol Data)

Estimator	Wilcoxon	Least Squares
Intercept	125.859 (4.62625)	246.880 (89.6751)
Gender Slope	-34.5544 (1.36563)	-10.6377 (26.4715)
Age Slope	3.1938 (.10989)	.2734 (2.1302)
$\hat{\rho}$	.1319	.0691
Scale	$\hat{\tau} = 5.8333$	$\hat{\sigma} = 10.4713$

all measurements collected on subject 19). The estimators for this reduced model were substantially different from the original estimators (see Table 11).

Table 11

Wilcoxon Estimates for Full Model Versus Model Excluding Subject 19

Estimator	Full Model	Excluding Subject 19
Intercept	119.301 (8.29945)	174.295 (29.1645)
Gender Slope	-35.8443 (2.44994)	-24.4925 (7.9622)
Age Slope	3.3273 (.19715)	1.6699 (.7287)
$\hat{\rho}$	.2336	.2941
Scale	$\hat{\tau} = 5.43807$	$\hat{\sigma} = 5.1807$

The robust measure described in Section 5.3 flags subject 19 as being influential,

$$WDBETA_{19} = 4.0776, \quad p = .0084.$$

However, it is not our recommendation to delete this subject but, instead, to further investigate the reason for its apparent abnormal values.

CHAPTER VII

CONCLUSION

The present thesis extends the theory of robust analyses for linear and nonlinear models to the Generalized Estimating Equations models.

Models can be linear or non-linear (e.g. Michaelis-Manten or Bateman relations for pharmacokinetic data) and computation can be handled by Gauss-Newton type algorithms.

The robust estimator, presented in Chapter III, based on the weighted Wilcoxon dispersion function, exists, is consistent and follows an asymptotically normal distribution. For the heteroscedasticity problem, where the errors are independent but have non-constant variances, we show that these robust estimates retain their consistency and asymptotic normality provided scale is consistently estimated. We have also investigated different covariance structures and provided estimation details for the AR(1) structure.

The asymptotic tests derived in Chapter IV were: the quadratic form-based Wald test for testing generalized linear model hypothesis, the gradient scores test and the drop in dispersion test. A test for different link functions is also derived.

In Chapter V, a check for outliers, based on the standardized residuals, was developed, using the asymptotic representation of the estimator. In addition, influential subjects are detected using a generalized Belsley's approach.

In Chapter VI real-data examples illustrated the estimator and the new testing techniques. The cholesterol example further proved the robustness of the Wilcoxon estimator when the data is stressed, in the presence of an outlier. The new method of detecting influential subjects is also put in practice with the cholesterol example, and one subject, suspected to be influential from the qqplots, is deemed to be influential by the WDBETA statistic.

The simulation results based on a multiple sclerosis example proved that the analyses based on the robust GEE are somewhat liberal. This is not surprising due to the liberalness of the Prais-Winsten procedure on linear model with time series errors. As with this simpler model, the liberalness seems to grow worse as ρ increases.

For the simpler model, McKnight et al. (1999) proposed a double bootstrap procedure which led to reasonable empirical α levels and empirical confidences, in general quite close to nominal. This procedure combines a Durbin two-stage estimation procedure with a bootstrap designed to estimate the bias in the estimate of the autoregressive parameters.

We intend to consider this bootstrap procedure in future research.

More research is needed in obtaining consistent robust estimates of the variance-covariance matrix, when it does not follow an AR(1) structure, but, for example, a compound symmetry or exchangeable structure. When these consistent estimates of the variance-covariance matrix are used, more research is needed in showing that the parameter estimators are consistent and follow an asymptotic normal distribution.

REFERENCES

Abebe, A. (2002). *Nonlinear Regression Based on Ranks,* doctoral dissertation, Western Michigan University, Kalamazoo, Michigan, USA.

Adimari, G., and Ventura, L. (2002). Robust Inference for Generalized Linear Models with Applications to Logistic Regression. *Statistics and Probability Letters.*

Belsley, D.A., Kuh, K., and Welsch, R.E. (1980). *Regression Diagnostics.* John Wiley and Sons, New York.

Breslow, N.E., and Lin, X. (1995). Bias correction in Generalized Linear Mixed Models with a Single Component of Dispersion. *Biometrika, 82,* 81-91.

Cantoni, E., and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *Journal of American Statistical Association, 96,* 1022-1030.

Chang, W.H., McKean, J.W., Naranjo, J.D., and Sheather, S.J. (1999). High-breakdown Rank Regression. *Journal of the American Statistical Association, 94,* 205-219.

Chi, E.M., and Reisel, G.C. (1989). Models for Longitudinal Data with Random Effects and AR(1) Errors. *Journal of American Statistical Association, 84,* 452-459.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15-18.

Crowder, M. (1995). On the use of a Working Correlation Matrix in Using Generalized Linear Models for Repeated Measures. *Biometrika, 82*, 407-410.

Diggle, P. J. (1988). An Approach to the Analysis of Repeated Measures. *Biometrics, 44*, 959-971.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1995). *Analysis of Longitudinal Data.* Oxford University Press, New York.

Dixon, S.L., and McKean, J.W. (1996). Rank-based Analysis of the Heteroscedastic Linear Model. *Journal of the American Statistical Association, 91*, 699-712.

D'yachkova, Y.D., Petkau, J., and White, R. (1997). Longitudinal Analysis for Magnetic Resonance Imaging Outcomes in Multiple Sclerosis Clinical Trials. *Journal of Biopharmaceutical Statistics, 7*, 501-531.

Gill, P.S. (2000). A Robust Mixed Linear Model Analysis for Longitudinal Data. *Statistics in Medicine, 19*, 975-987.

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

Hu, M., and Lachin, J.M. (2001). Application of Robust Estimating Equations to the Analysis of Quantitative Longitudinal Data. *Statistics in Medicine, 20*, 3411-3428.

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

1458.

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

Jones, R.H., and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics, 47*, 161-175.

Jorgensen, B. (1983). Maximum Likelihood Estimation and Large Sample Inference for Generalized Linear and Nonlinear Regresson Models. *Biometrika, 70*, 19-28.

Koul, H.L., Sievers, G.L., and McKean, J.W. (1987). An estimator of the scale parameters for the rank analysis of linear models under general score functions. *Scandinavian Journal of Statistics, 14*, 131-141.

Koul, H.L., and Saleh, A.K.M.E. (1993). R-estimation of the parameters of autoregressive [AR(p)] models. *Annals of Statistics, 21*, 534-551.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika, 73*, 13-22.

Liang, K.Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate Regression Analyses for Categorical Data (with Discussion). *Journal of the Royal Statistical Society, Series B, 54*, 3-40.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models.* 2nd edition, Chapman and Hall.

McKean, J.W., and Hettmansperger, T.P. (1976). Tests of hypotheses of the generalized linear model based on ranks. *Communications in Statistics. Thoery and Methods,5*, 693-709.

McKean, J.W., and Hettmansperger, T.P. (1978). A robust analysis of the generalized linear model based on one step R-estimates. *Biometrika, 65*, 571-579.

McKnight, S.D., McKean, J.W., and Huitema, B.E. (1999). A Double Bootstrap Method to Analyze Linear Models with Autoregressive Error Terms.

Miller, M.E., Davis, C.S, and Landis, J.R. (1993). The Analysis of Longitudinal Polytomous Data: GEE and Connections with Weighted Least Squares. *Biometrics, 49*, 1033-1044.

Morgenthaler, S. (1992). Least-absolute-deviations fits for Generalized Linear Models. *Biometrika, 79*, 747-754.

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

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*, 370-384.

Petkau, J., and White, R. (1995). *Longitudinal Analysis for the UBC 6-weekly frequent MRI substudy of the Betaseron Multiple Sclerosis Clinical Trial.* Biostatistics Research Report No 10, Department of Statistics, The University of British Columbia, Vancouver.

Preisser, J.S., and Qaqish, B.F. (1999). Robust Regression for Clustered

Data with Application to Binary Responses. *Biometrics, 55*, 574-579.

Prentice, R.L., and Zhao, L.P. (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrika, 47*, 825-839.

Rosner, B., Munoz, A., Tager, I., Speizer, F.E., Weiss, S.T. (1985). Use of a generalized autoregressive models for the analysis of longitudinal data. *Statistics in Medicine,4*, 457-467.

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

Sutradhar, B.C., and Das, K. (1999). On the Efficiency of Regresson Estimators in Generalized Linear Models for Longitudinal Data. *Biometrika, 86*, 459-465.

Terpstra, J.T., McKean, J.W., and Naranjo, J.D. (2000). Highly efficient Weighed Wilcoxon Estimates for Autoregression. *Statistics, 35*, 45-80.

Terpstra, J.T., McKean, J.W., and Naranjo, J.D. (2001). GR-estimates for an Autoregressive Time Series. *Statistics and Probability Letters, 51*, 165-172.

Wedderburn, R.W. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika, 61*, 439-447.

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

Zeger, S.L., and Liang, K.Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics, 42*, 121-130.

Zhang, D., and Davidian, M. (2001). Linear Mixed Models eith Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics, 57*, 795-802.

Zheng, B. (2000). Summarizing the Goodness of Fit of Generalized Linear Models for Longitudinal Data. *Statistics in Medicine, 19*, 1265-1275.