REGRESSION DIAGNOSTICS AND RESISTANT FITS FOR GENERALIZED ESTIMATING EQUATIONS

by
John S. Preisser
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2147

May 1995

REGRESSION DIAGNOSTICS AND RESISTANT FITS FOR GENERALIZED ESTIMATING EQUATIONS

by

John S. Preisser

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1995

Approved by:

Bahjat Qaqish Advisor

Reader

Gary Koch Reader

JOHN S. PREISSER Regression Diagnostics and Resistant Fits for Generalized Estimating Equations (Under the direction of Bahjat F. Qaqish.)

ABSTRACT

The Generalized Estimating Equations (GEE) procedure of Liang and Zeger (1986) can be highly influenced by the presence of unusual data points. Deletion diagnostics are introduced that consider leverage and residuals to measure the influence of a subset of observations on the estimated regression parameters and on the estimated values of the linear predictor. Computational formulae are provided which correspond to the influence of a single observation and of an entire cluster of correlated observations. The proposed diagnostics are generalizations of DBETA (Belsley et al. (1980)) and Cook's D (Cook (1977)) of linear regression. As an alternative approach, the influence of observations is addressed by Resistant Generalized Estimating Equations (REGEE). A generalization of the GEE procedure, REGEE gives parameter estimates and fitted values which are resistant to influential data. Robustness is achieved in REGEE through the inclusion of a diagonal weight matrix for each cluster in the estimating equations which downweight the multivariate response vector element-wise. The weights are defined with respect to leverage, corresponding to the Mallows class (Carroll and Pederson (1993)), or residual, the Schweppe class (Pregibon (1982)). The large sample and small sample properties are studied. An example of medical practice data is given to illustrate the use of the deletion diagnostics and REGEE for correlated binary regression.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Bahjat Qaqish, for his enthusiasm, patience and dedication in guiding me in this work. Thanks also to Dr. Gary Koch for his mentorship and for providing me with a great environment in which to work in the Biometric Consulting Laboratory. Special thanks go to Dr. Philip Sloane and colleagues in the Department of Family Medicine for providing me an opportunity to contribute and learn about gerontological research in an atmosphere of scholarship and friendship. Thanks also to Dr. P.K. Sen, Dr. Paul Stewart and Dr. Berton Kaplan, who complete my committee. I am grateful to the North Carolina Early Cancer Detection Program and the Lineberger Comprehensive Cancer Center for allowing the use of their data. The North Carolina Early Cancer Detection Program is funded by the National Cancer Institute. My efforts were partially supported by a National Institute of Environmental Health Sciences Training Grant.

This dissertation would not have been possible without the support of many people. I thank my father for his sacrifices in providing me with an undergraduate education and my mother for her prayers and love. Special recognition is given to the Newman Student Center at UNC and all my friends both there and in the Department of Biostatistics for there friendship and support. I especially thank Lisa Carmichael for her encouragement and support in the last months.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARE	Asymptotic Relative Efficiency
exp	exponential
DBETAC	approximate difference in beta due to deleting a cluster
DBETACS	standardized approximate difference in beta due to deleting a cluster
DBETAO	approximate difference in beta due to deleting an observation
DCLS	Cook's Distance statistic for a cluster
DOBS	Cook's Distance statistic for an observation
Eff_β	Finite Sample Efficiency
GEE	Generalized Estimating Equations
IRLS	Iteratively reweighted least squares
MLE	Maximum likelihood estimate
MSE	Mean Squared Error
RATIO	measure of efficiency loss in REGEE
REGEE	Resistant Generalized Estimating Equations

CHAPTER I

INTRODUCTION

The Generalized Estimating Equations (GEE) procedure of Liang and Zeger (1986) has been applied to a wide range of medical and biological applications in which the responses are correlated. Much of the appeal results from the broad capabilities of GEE, including modelling correlated binary outcomes and covariates that vary within clusters. Despite the frequent use of GEE, however, the influence of observations on parameter estimates and fitted values has not been addressed. This work proposes methodology to consider influence in multivariate regression models for discrete and continuous responses. The first part of this work introduces regression diagnostics for GEE in order to identify subsets of the data having a disproportionate effect on regression parameter estimates and fitted values. The second part introduces a modification of the GEE procedure called Resistant Generalized Estimating Equations (REGEE) which downweights observations based upon their influence.

The class of multivariate regression models considered by GEE and REGEE are models for marginal means, as distinguished from random effects models and conditional models. The difference lies in the interpretation of regression model parameters (Zeger et al. (1988)). In marginal models, the regression parameters represent the effect of predictors on the outcome, averaged over the population. These population-averaged models are commonly used in epidemiology, medicine and environmental science, to quantify risk factors or covariates with respect to a population. In random effects (or mixed effects) models, however, regression parameters have a subject-specific interpretation. Heterogeneity in regression parameters is explicitly modelled by allowing

the effect of a covariate to vary from subject to subject according to a distribution (often the normal distribution) specified up to a few unknown parameters. Finally, in conditional models, the estimated regression parameters have interpretations which are conditional on the values of the other outcomes (Zeger and Qaqish (1988); Rosner (1989)). These models have applicability in clinical settings, for example, where interest may be in the probability of disease in an individual given the disease status of other individuals in the family. An overview of methods for the analysis of longitudinal data is given by Zeger and Liang (1992).

This work is motivated by the lack of robustness of the GEE procedure. In particular, regression parameter estimates from an analysis based on GEE may be highly influenced by a small subset of the data. In other words, GEE lacks robustness in the sense of Huber (1964) who applied "robust" to statistical methods which have good properties when the data are contaminated by the presence of a few outliers or when the shape of the distribution deviates slightly from the assumed model. Our use of the word "robust" is not to be confused with that of Box (1953) who showed that the F-test for equality of variances behaves poorly in the presence of nonnormality. In his usage, "robust" characterized statistical methods that work well when there is some deviation from distributional assumptions, like normality or statistical independence. McCullagh and Nelder (1989, ch. 12) define these two different types of model departures, as isolated and systematic, respectively.

In order to identify isolated data points that depart from the specified model, regression diagnostics for GEE are developed that are generalizations of the diagnostics DBETA (Belsley, Kuh, and Welsch (1980)) and Cook's distance (Cook (1977)) for linear regression. These deletion diagnostics give the approximate change in regression parameters and fits when a single observation or cluster is omitted. Furthermore, they are obtained at little additional cost at convergence of the iteratively reweighted least squares algorithm of GEE. These diagnostics are discussed and applied to real data in

Chapter 3. The use of diagnostics, however, does not gaurentee that influential data will be identified, especially, for more complex multivariate models. Furthermore, outliers for binary regression are harder to identify than outliers for continuous data.

Robust regression procedures often identify influential data overlooked by deletion diagnostics. A robust multivariate regression procedure, REGEE, is proposed for continuous and discrete responses, and unlike GEE, it gives parameter estimates that are resistant to the influence of outliers and high leverage values. The REGEE procedure requires all of the model assumptions of GEE, but only that they hold for a majority of the data. REGEE is not a nonparametric procedure because the semi-parametric assumptions of GEE are not replaced.

Like GEE, REGEE applies to marginal models that have vectors of response variables corresponding to clusters whose components are correlated (usually positively) with one another. Dependency among observations occurs only within clusters, whereas responses from different clusters are assumed to be statistically independent of one another. Three designs, in particular, are among those used in research practice. In the first design, the response vector corresponds to repeated measures taken on subjects and parameter estimates are obtained for the marginal mean in a single regression model. When the responses are entirely different, separate regression models for the mean are in order, yet taking the correlation among responses into account through a multivariate procedure may provide more efficient estimation of regression parameters. Lastly, the response vector may correspond to a cluster of distinct subjects, such as a medical practice, and each component in the vector corresponds to a subject, such as a patient in the practice. Then, a single regression model often relates covariates to the response.

Regression parameter estimates from REGEE are extensions of robust estimates for generalized linear models. They are determined by an iteratively reweighted least squares (IRLS) algorithm that automatically downweights influential data. In the IRLS algorithm, each observation receives a weight between 0 and 1 corresponding to the

amount of its influence on parameter estimates. Weights based on an observation's leverage correspond to robust estimators of the Mallows Class (Carroll and Pederson (1993)). Weights depending on the value of the response, as well, give M-estimators (Hampel et al. (1986), Singer and Sen (1985)) of the Schweppe Class. Through the weights, REGEE balances the desire for robust protection and high efficiency.

In summary, REGEE provides resistant fits, in the sense of Huber (1964), that are not sensitive to large changes in a few observations (Pregibon, 1982). Chapter 4 introduces REGEE, presents its large sample properties, and demonstrates its resistance to influential clusters in an application of medical practice data. In chapter 5, the superior performance of REGEE over GEE in terms of robustness is shown through computer simulations by repeated sampling from a contaminated mixture model for binary responses (Copas, 1988). In particular, REGEE has smaller bias than GEE when data has small amounts of contamination. Since REGEE is proposed as an alternative methodology to GEE, the efficiency of REGEE with respect to GEE is evaluated analytically and at finite sample sizes through simulations from a model without contamination. Although REGEE is shown to be generally less efficient than GEE, for many practical situations its efficiency is high.

CHAPTER II

LITERATURE REVIEW

2.1 Dealing with Influence: Case-deletion diagnostics versus robust regression

2.1.1 Introduction

There is widespread agreement that concern for influential observations should be part of any analysis (Cook, 1986). In regression analysis, the traditional and most widely used approach in dealing with influence is to calculate deletion diagnostics which give the change (or approximate change) in the estimated regression parameters when a subset of the observations are deleted. Because the number of possible subsets is enormous, evaluation of these diagnostics is usually limited to the deletion of a single observation, or in regression for dependent responses, deletion of a cluster. An alternative way of dealing with influence is robust regression as developed by Hampel et al. (1986) for linear regression. In this approach, influential observations are automatically downweighted in the estimating equations to give resistant regression parameter estimates. In this chapter, deletion diagnostics and robust regession are reviewed for the univariate linear model, generalized linear models, and the multivariate linear model, respectively.

2.1.2 The univariate linear model

Consider the linear model. Let Y_i, $i = 1, ..., n$ be a sequence of i.i.d. random variables such that

$$Y_i = X_i\beta + \epsilon_i \quad i = 1, ..., n \tag{2.1}$$

where Y_i is the i-th observation, X_i is the ith row of the design matrix X, β is a p-vector of unknown parameters ($p \geq 1$), and ϵ_i is the i-th error. We suppose that ϵ_i has a symmetric distribution G (ϵ/σ) where $\sigma > 0$ is a scale parameter. The least squares

solution is $\hat{\beta} = (X'X)^{-1}X'Y$, the vector of fitted values is given by $\hat{\mu} = HY$ where $H = X(X'X)^{-1}X'$ is the hat matrix, and the residual vector is $\mathrm{E} = Y - \hat{\mu} = (I - H)Y$. An outlier is an observation with a large residual, defined $e_i = y_i - \hat{\mu}_i$. The leverage of an observation is $h_i = X_i(X'X)^{-1}X_i'$, the i-th diagonal element of H. A rule of thumb is to use $h_i > 2p/n$ to indicate points of high leverage.

There are various ways to standardize e_i. The i-th studentized residual is

$$r_i' = e_i/[s(1 - h_i)^{1/2}] \tag{2.2}$$

where s^2 is the residual variance or mean squared error (MSE). Another version of the expression above, called the deletion residual by McCullagh and Nelder (1989), replaces s with $s_{[i]}$ where the residual variance is calculated after deleting the i-th observation. Cook and Weisberg (1982) call these the internally studentized residual, and the externally studentized residual, respectively.

It is well known that the least squares estimate may be highly influenced by unusual observations. Accordingly, deletion diagnostics, routinely produced by common statistical computer packages, are used to identify such observations. Let $[i]$ denote an estimate evaluated without the i-th observation. If the i-th observation is deleted from the data set, the difference in the estimated regression parameter is given by

$$DBETA_i = \widehat{\beta} - \widehat{\beta}_{[i]} = (X'X)^{-1}X_i'(\frac{e_i}{1 - h_i}). \tag{2.3}$$

Belsley et al. (1980, p.13) recommend a scaled measure for the j-th element of β,

$$DBETAS_{ij} = DBETA_{ij}/[s_{[i]}^2(X'X)^{-1}]_{jj}^{1/2} \tag{2.4}$$

They give the change in the regression fit as

$$DFFIT_i = X_i(\widehat{\beta} - \widehat{\beta}_{[i]}) = \frac{h_i e_i}{1 - h_i} \tag{2.5}$$

For multiple case deletion, where m denotes the set of observations to be deleted,

$$DBETA_m = \widehat{\beta} - \widehat{\beta}_{[m]} = (X'X)^{-1}X'_m(I - H_m)^{-1}e_m \tag{2.6}$$

where $H_m = X_m(X'X)^{-1}X'_m$ and $e_m = y - X_m\widehat{\beta}$ (Atkinson, 1985, p. 20). The change in fit can be given as in (2.5), but it is often desirable to reduce this quantity to a scalar summary measure. In general, a class of norms is given by

$$D_m = (\widehat{\beta} - \widehat{\beta}_{[m]})'M(\widehat{\beta} - \widehat{\beta}_{[m]})/c.$$

Belsley et al. (1980, p. 32), propose a measure defined by $M = (X'_{[m]}X_{[m]})$ and $c = 1$, which reduces to

$$MDFFIT = e_m(I - H_m)^{-1}H_m e_m. \tag{2.7}$$

Cook (1986), however, states that $MDFFIT$ is a measure of influence on $\widehat{\beta}$ *and* the covariance matrix of $\widehat{\beta}$, and argues that if interest is in the influence on $\widehat{\beta}$ only, then one should use $M = X'X$. If, in addition, $c = ps^2$, the multiple case deletion Cook's distance (Cook and Weisberg, 1982) is obtained as

$$e_m(I - H_m)^{-1}H_m(I - H_m)^{-1}e_m/(ps^2). \tag{2.8}$$

For the deletion of a single observation, this is equivalent to the original Cook's distance (Cook, 1977; Cook, 1979) which is $h_i r_i'^2/[p(1 - h_i)]$. Most authors (Cook and Weisberg, 1982; Kleinbaum et al., 1988, p. 201) indicate that a value of about 1.0 would generally be considered large.

Robust Regression

In addition to identifying which observations are influential, robust regression automatically downweights observations in the estimation of $\hat{\beta}$. For the linear model, the standard approach is based on M-estimation as described by Hampel et al. (1986). A review of basic principles of M-estimation and robust linear regression follows.

Let $T(F)$ be a functional defined on a distribution function F (defined by a single parameter, θ), and associated with a function $\psi(Y,t)$, such that $T(F)$ is the solution t_o to the equation

$$\int \psi(Y,t_o)dF(Y) = 0. \tag{2.9}$$

Serfling (1980; chapter 7) calls $T(\cdot)$ the M-functional corresponding to ψ. For a sample $y_1, \ldots, y_n$ from F, the M-estimate $T(F_n)$ corresponding to ψ is the solution T_n of

$$\sum_{i=1}^{n} \psi(y_i, T_n) = 0. \tag{2.10}$$

Typically, (2.9) corresponds to minimization of some quantity

$$\int \rho(Y, t_o)dF(Y),$$

the function ψ determined by $\psi(Y, \theta) = \frac{\partial}{\partial\theta}\rho(Y, \theta)$ for $\rho(Y, \cdot)$ sufficiently smooth. If $Y_1, \ldots, Y_n$ are i.i.d. with density $f(Y)$, then $\rho = -\ln f(Y)$ gives the maximum likelihood estimator. The name "M-estimator" (Huber, 1964) comes from "generalized *m*aximum likelihood." The location problem is a special case of the single parameter model in which ψ functions of the type

$$\psi(Y, \theta) = \psi(Y - \theta)$$

are assumed in which $\int \psi\, dF = 0$ in order that T be Fisher consistent (i.e., $T(F_\theta) = \theta$ for all θ in Θ).

Hampel introduced the influence function to characterize the robustness properties of an estimator. Heuristically, it describes the effect of an infinitesimal contamination at the point y on the estimate, standardized by the mass of contamination. In other words, it measures the asymptotic bias caused by contamination in the observations. The influence function of T at F is

$$IF(Y;\psi,F) = \frac{\psi(Y, T(F))}{-\int (\frac{\partial}{\partial\theta})\,[\psi(Y, \theta)]_{T(F)}\, dF(Y)} = \frac{\psi(Y, \theta)}{\int\ \psi(Y, \theta)s(Y,\theta)\, dF(Y)}, \tag{2.11}$$

where $s(Y, \theta)$ is the maximum likelihood score function. Note that the denominator is (2.11) is equal to the Covariance of ψ and s evaluated at F. The asymptotic variance of the M-estimator is:

$$V(T, F) = E_F[\, IF^2(Y; \psi, F)] = \frac{\int \psi^2(Y, T(F))\, dF(Y)}{[\int (\frac{\partial}{\partial\theta})\, [\psi\,(Y, \theta)]_{T(F)}\, dF(Y)]^2} \tag{2.12}$$

Carroll and Ruppert (1988, chapter 7) give a heuristic argument for the asymptotic normality of M-estimates. They show, under certain regularity conditions,

$$N^{1/2}(T_n - \theta) \xrightarrow{D} N(0, V(T,F)) \tag{2.13}$$

Robustness theory gives consideration not only to the distribution of estimators under the assumed model, as in classical theory, but also under other probability distributions. Results (2.11), (2.12), and (2.13) can be evaluated at a neighboring distribution $F_{\theta^*} = F_{\theta^*}(y)$.

Assuming the model (2.1) has normally distributed errors, Huber (1973) proposed M-estimates, T_n, for β, defined by minimizing

$$\sum_{i=1}^{n} \rho((y_i - X_i'\beta)\, /\sigma)$$

which is equivalent to solving the vector equation

$$\sum_{i=1}^{n} \psi((y_i - X_i'T_n)\, /\sigma) X_i = 0. \tag{2.14}$$

In particular, he suggested

$$\psi_c\,(e_i) = e_i \cdot \min(1, c/|e_i|) \tag{2.15}$$

for some constant c. Combining (2.14) and (2.15) give weighted least squares estimate with weights of the form

$$g_i(e_i) = \min\,\{1,\, c/|e_i|\}, \tag{2.16}$$

which is equivalent to solving the estimating equations,

$$\sum_{i=1}^{n} g_i(e_i)X_i(y_i - X_i'T_n) = 0. \tag{2.17}$$

Hampel shows that the Huber estimate (2.16) bounds the "influence of residual" but not the "influence of position" of X_i in the factor space. Generalizing (2.14), in order to achieve greater robustness with respect to leverage, an M-estimator T_n for linear models is defined implicitly by the (vector) equation

$$\sum_{i=1}^{n} \eta\,(X_i, (y_i - X_i'T_n)/\sigma)X_i = 0, \tag{2.18}$$

where η satisfys certain regularity conditions concerning continuity and differentiability. The vector influence function of T at a distribution F is given by

$$\mathrm{IF}\,(Y; X, T, F) = \eta(X, Y - X' \cdot T(F)) \cdot M^{-1}(\eta, F)X \tag{2.19}$$

where

$$M(\eta, F) = \int \partial/\partial\beta\, \eta\,(X, Y - X' \cdot \beta)_{T(F)}\, XX'\, dF\,(Y; X).$$

These estimates are consistent and asymptotically normal with asymptotic covariance matrix

$$V(T,F) = \int IF\,(X, Y; T, F)\, IF'(X, Y; T, F)\, dF\,(Y; X) \tag{2.20}$$

$$= M^{-1}(\eta, F)\left\{ \int \eta^2\,(X, Y - X' \cdot T(F))\, XX'\, dF\,(Y; X) \right\} M^{-1}(\eta, F).$$

Hampel et al. (1986) define the gross-error sensitivity which measures the worst (approximate) influence which a small amount of contamination of fixed size can have on the value of the estimator. The unstandardized sensitivity is defined as

$$\gamma_u^*(\eta, F_\beta) = \sup_Y |\, \eta\,(X, Y - X'\beta)\,| \cdot \|M^{-1}X\|. \tag{2.21}$$

Schweppe type M-estimators are solutions to

$$\sum_{i=1}^{n} \psi\,[e_i/w_i(X_i)]\; w_i(X_i)X_i = 0 \tag{2.22}$$

which downweight leverage points only if the corresponding residual is large. Equation (2.22), which is a subclass of (2.18), is equivalent to

$$\sum_{i=1}^{n} g_i(e_i, X_i)\, w_i(X_i)(y_i - X_i' T_n). \tag{2.23}$$

Hampel et al. (1986) obtain an optimal M-estimator for the linear model, within the class (2.18), by minimizing the trace of the asymptotic covariance matrix (2.20) under the condition of a bound on the sensitivity (2.21). The solution is an η-function of Schweppe form

$$\eta\,(X_i, e_i) = w(X_i)\, \psi_c(e_i / w(X_i))$$

with certain weights, w, given implicitly by a formula found in Hampel et al. (1986).

In summary, optimal M-estimates for the linear model have been found by defining a class of estimators and imposing a bound on the influence that any given observation can have, while minimizing the variance of the estimator. Different classes of estimating equations with different definitions of sensitivity, and possibly different ways of minimizing the variance may give different "optimal" estimators.

2.1.3 Generalized Linear Models

A review of generalized linear models leads into a discussion of quasi-likelihood in which Generalized Estimating Equations (GEE) has its origins. Also, a review of resistant methods for generalized linear models suggests how to modify the GEE equations to make them more resistant to influential data, thus leading to the introduction of REGEE in chapter 4.

The theory of generalized linear models as proposed by Nelder and Wedderburn (1972) and expounded by McCullagh and Nelder (1989) provides a unifying theory for a general class of models which includes logistic regression, Poisson regression, and multiple linear regression, to name a few. The unifying component of these models is

that the distributions of the responses are of the exponential family. In particular, Y_i, the response for the i-th observation, has a distribution taking the form,

$$f(Y_i;\theta_i,\phi) = \exp\{(Y_i\theta_i - b(\theta_i))/a(\phi) + c(Y_i,\phi)\}. \tag{2.24}$$

The log-likelihood function is $\sum \log f(Y_i;\theta_i,\phi)$, and by differentiation with respect to θ_i it can be shown that $E(Y_i) = \mu_i = b'(\theta_i)$ and by further differentiation $\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi) = V(\mu_i)a(\phi)$, where θ_i is the canonical parameter, $V(\mu_i)$ is the variance function, and ϕ is the dispersion parameter. For the (scaled) Binomial distribution, B(m,μ)/m, with range 0, 1/m, 2/m, ... 1:

$$a(\phi) = 1/\mathrm{m}, \quad b(\theta_i) = \log(1 + \mathrm{e}^{\theta_i}) \text{ and } c(Y;\phi) = \log\binom{m}{mY}.$$

The mean is related to the canonical parameter by $\mu_i = \exp(\theta_i)/(1 + \exp(\theta_i))$, and the variance function, denoted by $V(\mu_i)$ because it depends on θ_i only through the mean, is equal to $\mu_i(1 - \mu_i)$.

A generalized linear model has three components: (1) the random component which is described above; (2) the systematic component or the linear predictor η given by $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, where x_{i1}, x_{i2}, ..., x_{ip} are the covariates for the ith observation; and (3) the link function, $g(\mu_i)$, which links the random and systematic components by the equation $\eta_i = g(\mu_i)$. The canonical link occurs when $\theta_i = \eta_i$ as for the Binomial response with logit link, $g(\mu_i) = \log\frac{\mu_i}{1-\mu_i}$, corresponding to logistic regression. The maximum likelihood estimate $\hat{\beta}$ is found by an iteratively reweighted least squares algorithm (McCullagh & Nelder, 1989, ch. 2, p. 40-43).

Model selection, i.e. choosing which set of covariates best describe Y, is accomplished through a measure of discrepancy called the deviance. The deviance compares the log likelihood of the model under consideration to the maximum

achievable value for the log-likelihood corresponding to a full model with n parameters, one for every observation. The deviance, which is easily derived from (2.24), is

$$D(y;\hat{\mu}) = \Sigma 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\} \tag{2.25}$$

where $\hat{\theta}_i = \theta(\hat{\mu})$ and $\tilde{\theta} = \theta(Y)$ estimate the canonical parameters under the two models. For the Binomial distribution,

$$D(y;\hat{\mu}) = \Sigma 2\{y_i \log(y_i/\hat{\mu}_i) + (m - y_i)\log[(m - y_i)/(m - \hat{\mu}_i)]\}. \tag{2.26}$$

In the generalized linear model, maximum likelihood estimates of the β_j's in the linear predictor are obtained by iterative weighted least squares. The statistical computer package, GLIM, can routinely fit generalized linear models.

Residuals are used to explore the adequacy of the fit of the model. In generalized linear models, the two most commonly used residuals are the Pearson residual,

$$r_{pi} = e_i/[V(\hat{\mu}_i)\hat{\phi}]^{1/2}, \tag{2.27}$$

and the deviance residual,

$$r_{di} = \text{sign}(e_i)d_i^{1/2}$$

where d_i is the deviance component for the i-th observation in (2.25). The studentized form of the Pearson residual is

$$r'_{pi} = e_i/[\hat{\phi}V(\hat{\mu}_i)(1 - h_i)]^{1/2} \tag{2.28}$$

and the corresponding form of the deviance residual is

$$r'_{di} = r_{di}/[\hat{\phi}(1 - h_i)]^{1/2}$$

where h_i is the i-th diagonal element of $H = W^{1/2}X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W^{1/2}$ where $W = (\partial\mu/\partial\eta)^2/V(\mu)$.

Model checking includes inpecting data for systematic departures from the model and checks for isolated departures from the model. Residual plots can be used informally for looking for systematic departures such as a wrong choice of link function, wrong choice of variance function or wrong scale of one or more covariates (McCullagh

and Nelder (1989), chapter 12). More formally one can specify a family of link functions such as $g(\mu;\lambda) = \mu^{\lambda}$ and test for adequacy of the chosen link function by using extended quasi-likelihoods (Nelder and Pregibon (1987)). A similar test for the variance function exists.

Model checking for isolated departures from the model includes diagnostics for case-deletion. Williams (1987) gives the one-step approximation for generalized linear models as

$$\hat{\beta} - \hat{\beta}_{[i]} \approx (X'WX)^{-1} X_i' W_i^{1/2} (1 - h_i)^{-1/2} r_{pi}'. \tag{2.29}$$

Pregibon (1981) introduced (2.29) for the logistic regression model, in which case a simpler form is obtained because $W_i = \mathrm{V}(\mu_i)$ due to the canonical link. The Cook statistic for generalized linear models is given by McCullagh and Nelder (1989, p. 407) as

$$D_i = (\hat{\beta}_{[i]} - \hat{\beta})^{\mathrm{T}} XWX(\hat{\beta}_{[i]} - \hat{\beta})/(p\hat{\phi}), \tag{2.30}$$

and is equivalent to $\mathrm{D_i} = \phi^{-1} p^{-1} h_i (1 - h_i)^{-1} r_{pi}^2$ (Williams, 1987). Kay and Little (1986) and Johnson (1985) have applied it to data fit by logistic regression models in order to identify points having a large effect on the fitted probabilities.

Robust Regression

Robust regression methods for the non-normal case have emphasized logistic regression, although the results extend to generalized linear models. In applying robustness theory to binary outcomes, one should consider how the nature of outliers for binary data differs from continuous data (Copas (1988)). In linear regression an outlier is either an extreme value of the response or an observation with a large residual in absolute terms. In logistic regression, the response is either a 0 or a 1, so that an outlier can only be defined in terms of its residual. The residual will be large if a 0 was observed

but the predicted value was very close to 1 (a downlier), or if a 1 is observed but the predicted value is very close to 0 (an uplier). An outlier might originate in the following ways: (1) The true value of the response was misrecorded, or transposed from a 0 to a 1, or from a 1 to a 0; (2) The observed value of the response is generated by a different model (contamination); (3) The values of the predictors were misrecorded; or (4) a very rare event was observed and the data point is "good" in that it came from the model. Even if the last case is true, we would not generally want one observation to greatly influence our overall results.

For binary responses, leverage plays a large role in the influence of an observation on generalized linear model regression parameter estimates. In many experimental settings where the design is balanced with respect to the predictors, robust regression may not be necessary if the responses are binary (especially if the number of replicates is reasonably large). However, in observational studies, where unusual design points are common, points with high leverage may arise and their influence becomes a concern. Predicted values very near 0 or 1 typically occur when the covariate values occupy an extreme position in the design space. Thus, in robust logistic regression, the weights usually consider leverage (Carroll and Pederson, 1993). The problem is that measuring leverage in not always easy to do. Three data points, for example, with identical but extreme design points might not be identified by a particular weight function because of a masking effect. Thus a question which has been considered for robust logistic regression is whether to downweight observations according to leverage only or according to residuals as well.

Many of the ideas behind robust regression for the linear model, (2.1), apply to robust regression for generalized linear models. The primary robust estimates for the generalized linear model are solutions to estimating equations of the form

$$0 = \sum_{i=1}^{n} g(x_i, x_i'\beta, y_i)\, x_i\, (y_i - \mu_i\,(x_i'\beta) - c(x_i, x_i'\beta)). \tag{2.31}$$

If $g_i = 1$ and $c_i = 0$, then (2.31) are the maximum likelihood score equations for a generalized linear model with canonical link. In general, however, robust estimates can be obtained by downweighting observations according to their leverage, fitted value, or residual, through use of a weight, g_i, which takes values between 0 and 1. Carroll and Pederson (1993) call g, the influence function, although it is not the same as the influence function (2.19) of Hampel et al. (1986). Recall that in logistic regression, and more generally, for generalized linear models with canonical link functions, $\partial\theta_i/\partial\eta_i = 1$. For other models, $\partial\theta_i/\partial\eta_i = w_i(x_i'\beta)$, and (2.31) still applies with this term absorbed in g to avoid redundancy.

Carroll and Pederson (1993) review the different subclasses of robust estimators that have been proposed for logistic regression. In principle, any influence function, g, that is used for the linear regression problem can be used (Pregibon, 1982). Estimates that are obtained from $g_i(x_i, x_i'\beta)$ which may depend on the design and predicted values, but not the responses, belong to the Mallows class. Influence functions, $g_i = (x_i, x_i'\beta, y_i)$, that are functions of Pearson or deviance residuals, give robust estimates of the Schweppe class. For Mallows class, $c_i = 0$ in (2.31) because the estimating equations are unbiased. For the Schweppe class, a debiasing factor, $c_i \neq 0$, is required to reduce the bias in the estimation of β. Often c_i is chosen such that the expected value of the right hand side of (2.31) is 0. The robust estimating equations for the linear model, (2.23), are a special case of (2.31). The debiasing factor is not needed for the location and scale family of distributions such as the normal distribution, so it does not appear in (2.23). However, for most generalized linear models it is required (Morgenthaler, 1992).

There are two important subclasses of robust esimators in the Mallows class. If the weights depend only on the design, $w_i = w(x_i)$, observations with high leverage are

downweighted because they are considered dangerous (or potentially influential). Carroll and Pederson (1993) call this method `leverage downweighting'. In the method of `prediction downweighting', the weights, $w_i = w(x_i'\beta)$ depend only on the fitted values.

In the Mallows class, consistent solutions $\hat{\beta}$ to equation (2.31) are asymptotically normally distributed. Let

$$V_{nj}(\beta) = n^{-1}\sum_{i=1}^{n} w_i^j \, x_i \, x_i' \frac{\partial}{\partial \eta_i} \mu_i(x_i'\beta) \tag{2.32}$$

where $\eta_i = x_i'\beta$. Then, asymptotically as n $\rightarrow \infty$,

$$n^{1/2}\,(\hat{\beta}_n - \beta) \sim \text{normal}\left(0,\, V_{n1}^{-1}(\beta)V_{n2}(\beta)V_{n1}^{-1}(\beta)\right). \tag{2.33}$$

The covariance matrix of $\hat{\beta}_n$ can be estimated consistently by

$$n^{-1}V_{n1}^{-1}(\hat{\beta})V_{n2}(\hat{\beta})V_{n1}^{-1}(\hat{\beta}).$$

The Schweppe estimates of Pregibon (1982) downweight the deviances, d_i, ($d_i = -\log(1-|e_i|)$ for logistic regression) by Huber's loss function which specifies the influence function,

$$\begin{aligned} g(e) &= 1 && \text{if } |e| \le 1 - e^{1/2h} \\ g(e) &= [-\tfrac{1}{2}h/\log(1-|e|)]^{1/2} && \text{otherwise} \end{aligned} \tag{2.34}$$

where $e_i = y_i - \mu_i(x_i'\beta)$, and h is the tuning constant taken to be $(1.345)^2$. His estimates, however, are not consistent at the logistic model since he uses $c_i = 0$ in (2.31). Copas (1988) refers to the resulting bias as the "overprediction of resistant fits". In other words, resistant fits in the Schweppe class downweight observations with large residuals which can result in parameter estimates that are larger in magnitude than their corresponding maximum likelihood estimates. See, for example, the robust parameter estimates in Tables 1, 2, and 3 of Pregibon (1982). Pregibon (1988), however, argues that the behavior of estimates *near* the logistic model is also important, and that his estimates based on likelihood tapering, (2.34), although inconsistent at the logistic

model, have asymptotic bias at neighboring distributions that is less than that of the maximum likelihood estimate.

Copas (1988) proposes robust logistic regression based on a misclassification model. In this model h, the tuning constant, has an interpretation as the probability of misrecording the binary response. Let

$$\mathrm{p}^* = \Pr(Y = 1 | x) = (1 - h)\mathrm{p} + h(1 - \mathrm{p}) \tag{2.35}$$

where

$$\mathrm{p} = \mu(x'\beta) = \left\{1 + \exp(-x'\beta)\right\}^{-1}.$$

The MLE $\hat{\beta}$ is an M-estimator solving

$$0 = \sum_{i=1}^{n} w_i(x_i'\beta, h)\, x_i\, \{Y_i - \mathrm{p}^*\}$$

where

$$w_i(x_i'\beta, h) = (1 - 2h)\, \mathrm{p}\, (1 - \mathrm{p})\, [\mathrm{p}^*(1 - \mathrm{p}^*)]^{-1}.$$

Because $\hat{\beta}$ is not consistent for β at the logistic model ($h = 0$), Copas (1988, equation (27)) provides a bias corrected version appropriate for small h. These estimates, however, are only approximately consistent for the logistic model and have small bias only when h is small and the p_i's are not "too close to 0 or 1", the so-called `weak regression' case. Carroll and Pederson (1993), provide a modification of the misclassification estimate which is consistent, is a member of the Mallows class, and has an interpretable tuning constant. Lastly, Kunsch et al. (1989) provide consistent and robust estimates in the Schweppe class. Analytical expressions for their finite sample bias are not feasible, however, due to the complexity of the estimate (Carroll and Pederson, (1993)).

Carroll and Pederson (1993) compare the performances of various robust estimates for logistic regression They consider: (1) a Mallows leverage downweighting estimate; (2) the Schweppe estimate by Kunsch et al. (1989); (3) the corrected

misclassification estimate by Copas (1988) for small (h = .01) and larger (h = .04) values of the tuning constant; and finally (4) the consistent misclassification estimate for a given h (.01 or .04) of Carroll and Pederson (1993). They evaluate the performance of these estimates in three different data sets. Their results show that for a small tuning constant, h = .01, (3) and (4) fail to downweight prediction outliers sufficiently. For one data set which had an observation with extreme leverage but that was a moderate prediction outlier, h = .03 failed to give satisfactory results. For data with extreme leverage points (1) performed well. There are some situations however when the leverage is masked because of a number of points with the same covariates, even though they may be in the extreme of the design. For the leukemia data (Cook and Weisberg (1982)) which had one prediction outlier whose leverage was masked by two other design points with identical value, (1) performed poorly, but (3) and (4) performed well. Although (2) performed satisfactorily in all cases, the authors do not recommend it for all possible situations.

For the Schweppe class, the choice of weight functions must be given special consideration for discrete data. Simpson et al. (1987) recommend weights corresponding to smooth ψ-functions for binary data, otherwise the estimates may be unstable at the values of the ψ-function where the derivative does not exist. For example, Huber's weight (2.16) is not recommended.

2.1.4 Models with Dependent Responses

Consider a general model

$$Y = X\beta + \epsilon \tag{2.36}$$

where Y is an N-vector, X is a $N \times p$ matrix of covariates, and β is a p-vector, and ϵ has mean 0 and a general covariance matrix, V. Let $V_{[i]}$ be the matrix V with the i-th row and column removed. Write

$$V = \begin{pmatrix} V_i & V_{i[i]} \\ V_{[i]i} & V_{[i]} \end{pmatrix}.$$

The matrices $Y_{[i]}$ and $X_{[i]}$ are Y and X with the i-th row removed. Christensen, Pearson, and Johnson (1992), show that the

$$\hat{\beta} - \hat{\beta}_{[i]} = (X'V^{-1}X)^{-1}\widetilde{X}_i' \frac{(\widetilde{Y}_i - \widetilde{X}_i\beta)}{(s_i - \widetilde{h}_i)} \tag{2.37}$$

where

$$\widetilde{X}_i = X_i - V_{i[i]}V_{[i]}^{-1}X_{[i]}, \quad \widetilde{Y}_i = Y_i - V_{i[i]}V_{[i]}^{-1}Y_{[i]},$$

$$\widetilde{h}_i = \widetilde{X}_i(X'V^{-1}X)^{-1}\widetilde{X}_i', \quad \textit{and} \quad s_i = V_i - V_{i[i]}V_{[i]}^{-1}V_{[i]i}.$$

Under independence, (2.37) reduces to (2.3). They also extend Cook's distance to

$$(\hat{\beta}_{[i]} - \hat{\beta})^{\mathrm{T}} XWX(\hat{\beta}_{[i]} - \hat{\beta})/(p\hat{\phi}) \tag{2.38}$$
$$= \frac{(\widetilde{Y}_i - \widetilde{X}_i\beta)^2\widetilde{h}_i}{p(s_i - \widetilde{h}_i)^2}.$$

Other classes of influence diagnostics have been proposed for multivariate linear models. Barrett and Ling (1992) consider a class of case-deletion diagnostics that apply to multivariate linear models with equal cluster sizes, only time-dependent covariates and variance constant across clusters. They consider diagnostics for clusters only, and not for observations within clusters.

Robust regression

M-estimation has been developed for regression with dependent continuous responses by Singer and Sen (1985), Bai et al. (1992), and Huggins (1993). Under specific assumptions, and under their respective models they obtain M-estimates which are consistent and asymptotic normal.

The model of Singer and Sen (1985) is

$$Y_k = X_k\beta + \epsilon_k, \qquad k = 1,\ldots,n \tag{2.39}$$

where Y_k are independent p-vectors (for some $p \geq 1$), X_k is an r-vector of known constants, and $\beta = (\beta_1, \ldots, \beta_p)$ is a (r $\times$ p) matrix of unknown parameters, and the ϵ_k are independent and identically distributed random vectors with distribution function F, defined on R^p. Classical procedures assume that the d.f. F is multi-normal. Singer and Sen (1985) assume that F has mean vector 0 and positive-definite covariance matrix, and a symmetric density function such that $\partial f(\epsilon)/\partial \epsilon_j$ exists, $j = 1, \ldots, p$. They propose a coordinatewise M-estimation method which applies different weights to the components of each response vector. The coordinatewise M-estimator of β is the solution of

$$\Sigma\, x_{ki}\, \psi_j\, \{(y_{kj} - \hat{\beta}_j' X_k)/\sigma_j\} = 0, \qquad (i = 1, \ldots, r;\ j = 1, \ldots, p). \tag{2.40}$$

An estimator is used in place of $\sigma = (\sigma_1^2, \ldots, \sigma_p^2)'$ if σ is unknown. Note that $\psi_j(x) = x$ ($j = 1, \ldots, p$) corresponds to the ordinary multivariate least squares estimator. With certain conditions on ψ (see C1 of Singer and Sen) and regularity conditions (see A1-A3, B of Singer and Sen), asymptotic normality of the M-estimator of $\hat{\beta}^* = (\hat{\beta}_1', \ldots, \hat{\beta}_p')'$ is obtained.

Singer and Sen (1985) compare the coordinatewise method to a simultaneous method (Maronna (1976)) that requires the stronger assumption that F be elliptically symmetric. Furthermore, M-estimation by the simultaneous method requires estimation of the entire covariance matrix Σ (which is a nuisance parameter), whereas the coordinatewise method would require only the estimation of diagonal elements of Σ. Finally, the coordinatewise method allows one to choose different score functions for different components of the observation vector, whereas the simultaneous method does not. This added flexibility is useful when some components are expected to produce more outliers than others.

One disadvantage of the above methods described in Singer and Sen (1985), is that the model in (2.39) does not allow time independent covariates. Bai, Rao and Wu (1992) consider a special case of (2.36) where X_i and Y_i represent the covariate matrix

and response vector for the i-th cluster and $\epsilon' = (\epsilon_1', ..., \epsilon_k')$ where ϵ_i, $i = 1,..., k$, are i.i.d. n-vectors. This model allows for time independent covariates and has (2.39) as a special case. Bai et al. (1992) prove consistency and asymptotic normality of M-estimates defined by minimizing

$$\Sigma\, \rho(Y_i - X_i\beta)$$

for any convex function ρ of p variables under a minimal set of conditions on X_i, and the random error ϵ_i . They do not assume that ψ is continuous.

The robust methods of Singer and Sen, and Bai et al. are concerned with downweighting observations in the multivariate model with respect to residuals. Thus what is required for the simultaneous method of Maronna (1976), which is discussed by Singer and Sen (1985), is a function that reduces a vector of residuals to a scalar measure of distance. If one was to take leverage into account, what would be needed is a scalar measure of the difference between matrices. This is more complex than the straight forward extension of the component-wise idea of Singer and Sen to account for leverage in a multivariate model with time-varying covariates. By considering components in the response vector one at a time, a weight based on leverage may be defined by a score function which reduces the vector of covariates to a scalar weight for each component. Allowing for time-varying covariates, the component-wise method of Singer and Sen will be considered in extending GEE to REGEE in Chapter 4.

Huggins (1993) provides robust estimation of regression parameters in a multivariate normal repeated measures model by modifying the multinormal likelihood score equations. In his model, $Y_i \sim MVN_{n_i}(\mu_i, \Omega_i)$, $\Omega_i^{-1} = B_i^{-T} B_i$, $Z_i = B_i^{-1}(Y_i - \mu_i)$, and $D = \partial\mu/\partial\beta$. The score equations are

$$\sum D_i' B_i^{-T} Z_i.$$

He modifies them by applying a function, ψ, to obtain

$$\sum D_i' B_i^{-T} \psi(Z_i).$$

Whereas, Singer and Sen (1993) standardize the responses element-wise, Huggins (1993) standardizes the responses vector-wise. The drawback is that the estimates may be more sensitive to misspecification of the correlation structure. REGEE will give consistent estimation of β, even if the correlation is misspecified.

2.2 Estimating Function Theory

2.2.1 Quasi-likelihood

Quasi-likelihood (Wedderburn (1974); McCullagh (1983)) is a semi-parametric method of estimation in which the likelihood function is not fully specified. Rather, we specify the functional forms of the mean and the variance only. The scale parameter ϕ is estimated.

The notion of a quasi-likelihood comes from observing that the score equation corresponding to a generalized linear model depends only on the mean μ_i and the variance function $V(\mu_i)$, the value of ϕ not affecting the estimation of β.

$$\frac{\partial}{\partial \beta_j} \Sigma \log f(y_i;\theta_i,\phi) = \Sigma \frac{\partial}{\partial \beta_j} \{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i,\phi)\}$$

$$= \Sigma \frac{\partial}{\partial \theta_i} \{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i,\phi)\} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \qquad (2.41)$$

$$= \Sigma \frac{1}{a(\phi)} (y_i - b'(\theta_i))[V(\mu_i)]^{-1} \frac{\partial \mu_i}{\partial \beta_j}$$

For generalized linear models with canonical links $\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = 1 \times x_i = x_i$, so that (2.41) simplifies to the familiar form of the score equations for the linear model:

$$\Sigma(y_i - b'(\theta_i))x_i = 0.$$

The score equations (2.40) can be written in vector form as

$$U(\beta) = D'V^{-1}(Y - \mu)/\sigma^2 \tag{2.42}$$

where $D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$ and D is n $\times$ p, $V = V(\mu) = \text{diag}\{V_1(\mu_1), ..., V_n(\mu_n)\}$ if the responses are independent, and σ^2 is the scale parameter (previously denoted ϕ) assumed constant over the marginal responses.

In practice, we may not be able to specify the distribution that generates the responses, but we may be able to specify (a) how the mean response changes with the covariates and (b) how the variability of the response changes with the average response, and (c) whether the observations are statistically independent. From this information we determine U in (2.42) which may or may not correspond to a log likelihood for a known distribution. The integral,

$$Q(\mu; y) = \int_y^\mu \frac{y-t}{\sigma^2 V(t)}\, dt,$$

if it exists, is called a quasi-likelihood.

Because quasi-likelihood estimators usually do not have closed form expressions, but are found by iteration, we study their properties by studying their estimating functions. U has many properties in common with a log-likelihood derivative:

$$E(U) = 0, \tag{2.43}$$

$$\text{var}(U) = -E(\frac{\partial U}{\partial \beta}) = D'V^{-1}D/\sigma^2 = i_\beta \tag{2.44}$$

For quasi-likelihood functions, i_β plays the same role as the Fisher information for ordinary likelihood functions. Under certain regularity conditions, $\text{cov}(\hat{\beta}) \approx i_\beta^{-1}$, and $\hat{\beta} \sim N(\beta, i_\beta^{-1})$.

If the data are dependent, then the quasi-likelihood estimating equations require a modification. We assume that $\text{cov}(Y) = \sigma^2 V(\mu)$ where $V(\mu)$ is a symmetric positive-definite $n \times n$ matrix of known functions $V_{ij}(\mu)$, no longer diagonal. Equations (2.43) and (2.44) still hold, and consistency and asymptotic normally of $\hat{\beta}$ arise under certain regularity conditions. There is essentially no difference in the discussion of quasi-

likelihood for independent and dependent observations except for the following. If the score vector $U(\beta)$ is to be the gradient vector of a log likelihood or quasi-likelihood, it is necessary and sufficient that the derivative matrix of $U(\beta)$ with respect to β be symmetric. In general, however, this is not the case. The implications of this for inference is not entirely clear according to McCullagh and Nelder (1989).

Quasi-likelihood is commonly applied to data which exhibit over-dispersion. Over-dispersion is present when the variance of the response Y exceeds that expected under the exponential family model. For grouped binary data, the binomial distribution specifies that the variance of the response is given by $\mathrm{var}(Y_i) = m_i\pi_i(1-\pi_i)$, where m_i is the cluster size and π_i is the probability of a positive response. Under the method of quasi-likelihood, $\mathrm{var}(Y_i) = \phi m_i\pi_i(1-\pi_i)$. If $\phi > 1$, the data exhibit over-dispersion. Under-dispersion, represented by $\phi < 1$, is less common. This mathematical form of over-dispersion in which ϕ does not depend on m_i can be generated by the notion of clustering in the population. The dispersion parameter ϕ can be estimated from the residuals under the model which assumes the observations are binomially distributed (McCullagh and Nelder, section 4.5). An alternative model for over-dispersion is provided by the beta-binomial model, which specifies ϕ as a linear function of m_i rather than as a constant (Crowder (1978)). Liang and McCullagh (1993) present examples of binary dispersed data and give diagnostic tools for determining the more plausible model for over-dispersion. They conclude that neither model seems to be preponderant, but that for any given data set one model may be better than the other. Of five data sets they consider, they were unable to identify a prefered model for three of them. For the other two data sets in which discrimination was possible, the choice of model influenced regression inference.

Besides problems of over-dispersion, there are many other situations where quasi-likelihood might be applied. For example, Preisser, Koch and Shockley (1993) analyze data obtained from an experiment concerning tracheal reconstruction in rats

carried out by researchers in the Department of Surgery in the University of North Carolina School of Medicine. A quasi-likelihood analysis specifying the binomial variance function $V(\mu) = \mu(1-\mu)$ and identity link function was employed in order to obtain regression parameter and standard error estimates which corresponded to comparisons of the conditions under which surgery was performed.

2.2.2 Optimal estimating functions

Often, the estimators that are commonly used in statistical practice do not have closed form expressions. Their distributional properties are more easily ascertained by studying the properties of their corresponding estimating functions. Optimality theory of estimating functions suggests estimators which are, in some sense, best. We review this theory, not only to describe the properties of quasi-likelihood in a larger context, but also to motivate discussion in Chapter 3 of optimal robust estimators.

An estimating function is a function, g, of the data Y and a parameter of interest θ, that when set to 0, gives an estimating equation, $g(Y,\theta) = 0$, from which, upon solving, gives an estimator of θ. For a single parameter θ, Godambe (1960), showed that the score function corresponding to maximum likelihood estimation is in a certain sense, optimal, in a certain class, G, of "regular estimating functions". Specifically, he defines regularity by requiring the following for every $g \in G$: (i) $E[g(Y,\theta)|\theta] = 0$; (ii) $E(\partial g/\partial\theta|\theta) \neq 0$; and (iii) $\int g(y,\theta)p(y,\theta)dy$ differentiable under the integral sign, where $p(y,\theta)$ is the probability density function of the random variable Y. He defines the following optimality criterion: let $g_s = g/[E(\partial g/\partial\theta|\theta)]$ be the standardized version of g. Then the optimal estimating function in G, is g^* if $E(g_s^{*2}) \leq E(g_s^2)$ for all $g \in G$. He showed that $g^* = \partial \log p/\partial\theta$. This result is strongly related to the Cramer-Rao inequality which establishes the lower bound on the variance of an unbiased estimate.

Bhapkar (1972) and Morton (1981) present a generalization of Godambe (1960) to the multi-parameter case. Define the vector estimating equations $g(Y,\theta) = 0$, and

also define $V_g = \mathrm{cov}(g(Y,\theta))$ and $B_g = E_\theta[\partial g/\partial\theta]$. Let $M_g = B_g' V_g^{-1} B$, and like Morton let it be called the efficiency matrix of g (Bhapkar calls it the information in g). Then, g^* is optimal if $M_{g^*} - M_g$ is semi-positive definite for all g in a certain class. In the class of unbiased (vector) estimating functions, under regularity conditions analagous to Godambe (1960), the maximum likelihood (vector) score equation is optimal.

Godambe and Heyde (1987) distinguish between two different optimality criteria for the multi-parameter problem. One is based on fixed sample sizes like that of Morton (1981) and Godambe (1960). The other considers g^* as optimal if it minimizes the asymptotic covariance matrix of $\hat{\theta}$ which solves $g(Y,\theta) = 0$.

Morton (1981) considers optimal estimating functions for θ in the presence of nuisance parameters ζ when the likelihood is not fully known. He supposes, however, that a pivot $h(x,\theta)$ exists whose distribution depends on θ only. For a class of unbiased estimating functions which are linear in the pivot, taking the form $g = C'[h - E(h/\theta,\zeta)]$, the optimal estimating function is $g^* = B_h' V_h^{-1} h$. The choice of g is not necessarily obvious, yet it is worth noting that $g = Y - \mu$ and $\theta = \beta$ gives the quasi-score function (2.41).

In the theory of estimating functions, quasi-likelihood has certain optimality properites. It is well known that least squares is optimal within the class of linear unbiased estimators (Cox and Hinkley (1968)). The least squares solution $\hat{\beta}$ is better than any other $\tilde{\beta}$ in that $\mathrm{var}(a'\hat{\beta}) \leq \mathrm{var}(a'\tilde{\beta})$. This optimality holds for weighted least squares in an asymptotic sense. Because quasi-likelihood is essentially an extended theory of non-linear least squares, it inherits certain optimality properties (McCullagh (1991)). Quasi-likelihood solutions are optimal in the sense that in the class of linear unbiased estimating equations the asymptotic variance is minimized (Godambe and Heyde (1987)). If the underlying distribution comes from an exponential family which corresponds to the quasi-likelihood score equation, then the solutions will have full asymptotic efficiency (Firth (1987)). Unbiasedness is always a desired property for

estimating functions, although we sometimes settle for approximate unbiasedness. Linearity, on the other hand is somewhat restrictive. Crowder (1987) considers the larger class of quadratic estimating functions.

CHAPTER III

DELETION DIAGNOSTICS FOR GENERALIZED ESTIMATING EQUATIONS

3.1 Introduction

Despite the frequent use of the Generalized Estimating Equations procedure, diagnostics do not exist that can identify observations or clusters which have a disproportionately large influence on the estimated regression parameters. This chapter introduces deletion diagnostics which account for the leverage and residuals in a set of observations to determine their influence on regression parameters and fitted values. When that set consists of only one observation we call them "observation-deletion" diagnostics. When it consists of a whole cluster, we call them "cluster-deletion" diagnostics. There is widespread agreement that concern for influential observations should be part of any analysis (Cook, 1986). This paper introduces influence measures which will make an analysis based on Generalized Estimating Equations more complete. They are generalizations of diagnostics for dependent responses as found in Christensen, Pearson and Johnson (1992), and of diagnostics for generalized linear models (Pregibon, 1981; Williams, 1987; McCullagh and Nelder, 1989, chapter 12). They reduce to well known measures of influence in linear regression found in Cook and Weisberg (1982), Belsley et al. (1980), and in a review paper by Chatterjee and Hadi (1986). In section 2, we review Generalized Estimating Equations and introduce some notation. Computational formulae are provided in section 3 which are based on one-step approximations (Pregibon, 1981), while accounting for within-cluster correlation.

Interpretation of the proposed measures is discussed through an illustrative example in section 4.

3.2 Generalized Estimating Equations

We review Generalized Estimating Equations giving special attention to the factors in the model fitting process which are most closely associated with influence, namely, leverage and residual. Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ be a n_i-vector of response values for $i = 1, \ldots, K$, and $X_i = (X'_{i1}, \ldots, X'_{in_i})'$ is a $n_i \times p$ matrix of covariate values. Throughout the chapter, clusters are indexed by i and observations by t. We consider models where the forms of the first two moments for the marginal distribution of Y_{it} are

$$E(Y_{it}) = \mu_{it}, \qquad g(\mu_{it}) = \eta_{it} = X_{it}\beta, \qquad \mathrm{var}(Y_{it}) = f_{it}\phi. \tag{3.1}$$

In the terminology of generalized linear models, $g(\mu_{it})$ is the link function which determines the relationship of the mean with the linear predictor η_{it}, $f_{it} = f(\mu_{it})$ is the variance function, β is a $p \times 1$ vector of regression coefficients, and ϕ is the scale parameter, either known or to be estimated. Estimates of β are obtained by solving the Generalized Estimating Equations

$$\sum_{i=1}^{K} D'_i \{A_i R_i(\alpha) A_i\}^{-1} (Y_i - \mu_i) = 0, \tag{3.2}$$

where $D_i = \partial\mu_i/\partial\beta$ is a $n_i \times p$ matrix, $A_i = \mathrm{diag}(f_{it}^{1/2})$ is a $n_i \times n_i$ diagonal matrix, and $R_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix that depends on unknown parameter vector α. A solution is obtained by alternating between estimation of ϕ, α, and β, using method of moment estimators for ϕ and α. We define $N = \Sigma n_i$, the $N \times 1$ vector $Y = (Y'_1, \ldots, Y'_K)'$, the $N \times p$ matrix $X = (X'_1, \ldots, X'_K)'$ assumed to be of full column rank, and $D^* = \partial\eta/\partial\mu$, a $N \times N$ diagonal matrix with nonzero elements

$d_{it} = \partial\eta_{it}/\partial\mu_{it}$. Estimation of β is done with iteratively reweighted least squares by regressing the working response vector $Z = X\hat{\beta} + D^*(Y - \hat{\mu})$ on X with block diagonal weight matrix W whose i-th block, corresponding to the i-th cluster, is the $n_i \times n_i$ matrix

$$W_i = D^{*-1}_i A_i^{-1} R_i^{-1}(\hat{\alpha}) A_i^{-1} D_i^{*-1}, \text{ and } D_i^* = \text{diag}(d_{i1}, ..., d_{in_i}).$$

Liang and Zeger (1986) show that, under regularity conditions, as $K \to \infty$, $K^{1/2}(\hat{\beta} - \beta)$ is asymptotically multivariate Gaussian with mean vector 0 and covariance matrix given by

$$J_{\hat{\beta}} = \lim_{K\to\infty} K J_1^{-1} J_2 J_1^{-1}, \tag{3.3}$$

$$\text{where } J_1 = \sum_{i=1}^{K} D_i'\{A_i R_i(\alpha) A_i\}^{-1} D_i,$$

$$J_2 = \sum_{i=1}^{K} D_i'\{A_i R_i(\alpha) A_i\}^{-1} \text{cov}\,(Y_i)\{A_i R_i(\alpha) A_i\}^{-1} D_i.$$

The "robust" or sandwich variance estimate of $\hat{\beta}$ is obtained by replacing cov (Y_i) by $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$ and β, ϕ, and α by their estimates in (3.3). It is robust in the sense that it consistently estimates $J_{\hat{\beta}}$ even if $R(\alpha)$ is misspecified. If $R(\alpha)$ is correctly specified, cov $(Y_i) = A_i R_i(\alpha) A_i \phi$ and (3.3) reduces to $\phi \lim K J_1^{-1}$ and the respective "naive" variance estimator of $\hat{\beta}$ is

$$\phi(\sum_{i=1}^{K} X_i' W_i X_i)^{-1}.$$

A current estimate of β is updated by

$$\hat{\beta}_{\text{new}} = (X'WX)^{-1} X'WZ,$$

evaluating the right hand side at the current estimate. Then $\hat{\beta}_{\text{new}}$ is used to update $\hat{\eta} = X\hat{\beta}_{\text{new}} = HZ$, where $H = QW$ and $Q = X(X'WX)^{-1}X'$. The asymmetric and idempotent projection matrix, H, maps the current value of Z into estimated values of the linear predictor. The diagonal elements of H, denoted by h_{it}, correspond to the amount of leverage of the response on the corresponding fitted value. The average of

the h_{it} is p/N which follows from $\text{tr}(H) = p$. The leverage of a cluster is contained in the matrix $H_i = Q_i W_i$ where $Q_i = X_i(X'WX)^{-1}X_i'$, and can be summarized by $\text{tr}(H_i)$ which has the additive property of being the sum of observation leverages.

The estimated adjusted residual vector is

$$E = D^*(Y - \hat{\mu}) = Z - \hat{\eta} = (I - H)Z. \tag{3.4}$$

Considering H and the current estimate of β as non-random, the variances of E and $\hat{\eta}$ are easily obtained by observing that $\text{var}\,(Z) = D^*\text{var}\,(Y)D^*$. If the correct weights, $W = \phi\{\text{var}\,(Z)\}^{-1}$ are applied, $\text{var}\,(E) = \phi(W^{-1} - Q)$ and $\text{var}\,(\hat{\eta}) = \phi Q$; by "correct" we mean that the working correlation structure is correctly specified, and $\hat{\alpha} = \alpha$. Additionally, $\text{var}\,(E_i) = \phi(W_i^{-1} - Q_i)$ and $\text{var}\,(\hat{\eta}_i) = \phi Q_i$, where $E_i = D_i^*(Y_i - \hat{\mu}_i)$ is the estimated residual of the i-th cluster. In the next section, the matrices Q_i and W_i appear in formulae which are proposed for measuring the influence of a cluster on the regression parameter estimates.

3.3 Measures of influence based on case-deletion

We consider the effect of deleting one or more observations on $\hat{\beta}$. Exact formulae would require complete iteration for every subset of observations deleted. Clearly this is computationally prohibitive, even in relatively small data sets. We introduce computationally feasible one-step approximations, like those of Pregibon (1981) for generalized linear models. In addition to considering the change in individual $\hat{\beta}_j$'s, formulae are given for the effect of deletion of one or more observations on the estimated values of the linear predictor. In other words, we examine the effect of deletion on the $\hat{\beta}_j$'s simultaneously. Formulae are given for the change caused by deleting an arbitrary subset of observations, and for two useful special cases: deleting one observation and deleting one cluster.

3.3.1 Assessing the influence of case deletion on $\hat{\beta}$

Let m index the subset of m observations that are to be deleted, $[m]$ denote the remaining observations, and let $\hat{\beta}_{[m]}$ denote the regression parameter estimate when the set m is removed from the data. Without loss of generality, assume that the observations to be deleted are the first m components of Z, and let W be partitioned as

$$W = \begin{pmatrix} W_m & W_{m[m]} \\ W_{[m]m} & W_{[m]} \end{pmatrix}.$$

All vectors and matrices will be partitioned in a parallel manner. Also define $V = W^{-1}$.

THEOREM 1. *The one-step approximation for $\hat{\beta}_{[m]}$ is:*

$$\hat{\beta}_{[m]} \approx \hat{\beta} - (X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{E}_m,$$

where

$$\tilde{X}_m = X_m - V_{m[m]}V_{[m]}^{-1}X_{[m]}, \quad \tilde{Q}_m = \tilde{X}_m(X'WX)^{-1}\tilde{X}'_m,$$

$$\tilde{E}_m = \tilde{Z}_m - \tilde{X}_m\hat{\beta} = E_m - V_{m[m]}V_{[m]}^{-1}E_{[m]}.$$

$$\textit{and } \tilde{Z}_m = Z_m - V_{m[m]}V_{[m]}^{-1}Z_{[m]}.$$

The proof is in Appendix 1. The matrices, $\tilde{X}_m, W_m^{-1}, \tilde{Q}_m$, are the building blocks of multiple case-deletion diagnostics, (2.36), for the linear model with general covariance matrix, V, considered by Christensen, Pearson, and Johnson (1992). The vector $\tilde{E}_m$ is introduced in order to generalize their result to other link functions, giving a one-step approximation for the change in the regression parameter estimates obtained from (3.2). Recall that $E = (E'_m, E'_{[m]})'$ is the residual of the working response vector given by (3.4). The dimension of $V_{[m]}^{-1}$ is $N - m$ by $N - m$ so that THEOREM 1 is not computationally feasible, except for special cases, two of which are now examined.

First, a cluster-deletion diagnostic is introduced which measures the effect of a single cluster on the estimated regression parameter vector. Let the subscript $[i]$ denote estimates evaluated without the i-th cluster.

COROLLARY 1.1 *The one-step approximation for* $\hat{\beta} - \hat{\beta}_{[i]}$ *is*:

$$DBETAC_i = (X'WX)^{-1} X_i' \, (W_i^{-1} - Q_i)^{-1} E_i. \tag{3.5}$$

PROOF. Applying THEOREM 1, $V_{i[i]} = 0$ implies $\tilde{E}_i = E_i$, $\tilde{X}_i = X_i$ and $\tilde{Q}_i = Q_i$.

The simple structure of (3.5), resulting from the fact that W and V are block diagonal matrices, provides a formula which is computationally feasible. An alternative expression to (3.5) is given by observing that $(W_i^{-1} - Q_i)^{-1} = W_i(I - H_i)^{-1}$.

Next, we introduce an observation-deletion diagnostic. Let $\hat{\beta}_{[it]}$ denote estimates evaluated using all the data except the t-th observation of the i-th cluster, and let the subscript $i[t]$ denote matrices corresponding to the i-th cluster without the t-th observation. Let the matrices W_i and V_i be partitioned as

$$W_i = \begin{pmatrix} W_{it} & W_{it[t]} \\ W_{i[t]t} & W_{i[t]} \end{pmatrix}, \quad V_i = W_i^{-1} = \begin{pmatrix} V_{it} & V_{it[t]} \\ V_{i[t]t} & V_{i[t]} \end{pmatrix}.$$

COROLLARY 1.2 *Let* $E_{it} = D_{it}^*(Y_{it} - \hat{\mu}_{it})$ *and* $E_{i[t]} = D_{i[t]}^*(Y_{i[t]} - \hat{\mu}_{i[t]})$. *Then one-step approximation for* $\hat{\beta} - \hat{\beta}_{[it]}$ *is*:

$$DBETAO_{it} = (X'WX)^{-1} \, \tilde{X}_{it}' \, \frac{\tilde{E}_{it}}{W_{it}^{-1} - \tilde{Q}_{it}} \, , \tag{3.6}$$

where

$$\tilde{X}_{it} = X_{it} - V_{it[t]} V_{i[t]}^{-1} X_{i[t]}, \quad \tilde{Q}_{it} = \tilde{X}_{it}(X'WX)^{-1} \, \tilde{X}_{it}',$$

$$\textit{and} \quad \tilde{E}_{it} = E_{it} - V_{it[t]} V_{i[t]}^{-1} E_{i[t]}.$$

Note that W_{it}, $\widetilde{Q}_{it}$ and E_{it} are scalars.

PROOF. Applying THEOREM 1, $V_{[m]m} = \begin{bmatrix} V_{i[t]t} \\ 0 \end{bmatrix}$ and because V is block diagonal $V_{m[m]}V_{[m]}^{-1}X_{[m]} = V_{it[t]}V_{i[t]}^{-1}X_{i[t]}$, and $V_{m[m]}V_{[m]}^{-1}E_{[m]} = V_{it[t]}V_{i[t]}^{-1}E_{i[t]}$. The result follows.

The influence diagnostics in THEOREM 1, COROLLARY 1.1 and COROLLARY 1.2 are generalizations of a number of known results in which the estimating equations are special cases of (3.2). If independence is assumed, the distinction between one cluster and one observation vanishes as W becomes a diagonal matrix, and the result in either of the corollaries above reduces to the one-step approximation for generalized linear models given in section 3 of Williams (1987). Applying COROLLARY 1.1 with all cluster sizes equal to 1 yields (2.29).

Next consider an identity link, $g(\mu_{it}) = \mu_{it}$, for the model in (3.1) with a general covariance structure, V. The identity link implies $D^* = I$ which implies $\widetilde{Z}_m = \widetilde{Y}_m$, where

$$\widetilde{Y}_m = Y_m - V_{m[m]}V_{[m]}^{-1}Y_{[m]}.$$

A formula identical to that of THEOREM 1 is obtained, but with $W = V^{-1}$and $\widetilde{E}_m = \widetilde{Y}_m - \widetilde{X}_m\hat{\beta}$. For observation-deletion, (2.36) which is proposition 3 of Christensen, Pearson and Johnson (1992) is obtained as a special case of COROLLARY 1.2.

Consideration of independence, identity link, and constant variance simultaneously give well known results for the multiple linear regression model. In this case, $W = I$, and $\widetilde{Q}_m = X_m(X'X)^{-1}X_m'$. For multiple case-deletion, THEOREM 1 reduces to (2.6) and for observation-deletion it reduces to (2.3).

The proposed diagnostics can be standardized using the variances of $\hat{\beta}$ based on the complete data or with the subset m omitted from the calculation. For example, the

standardized one-step approximation for the change in $\hat{\beta}_j$ due to the deletion of the i-th cluster is

$$DBETACS_{ij} = DBETAC_{ij} / \{\hat{\phi}\,(X'WX)^{-1}\}_{jj}^{1/2}. \tag{3.7}$$

For single-case deletion in multiple linear regression this measure is equal to (2.4) if ϕ is estimated by $s^2_{[i]}$ which is the usual estimate of variance calculated without the i-th case. Notice that standardization in (8) is achieved by dividing by the naive standard error of the j-th coefficient based on all the data. Alternatively, the studentized approximation for the change in β_j is obtained by dividing by the square root of $\{\hat{\phi}(X_{[i]}W_{[i]}X_{[i]})^{-1}\}_{jj}$ which omits the effect of the i-th cluster on the standard error of $\hat{\beta}_j$. Both of theses scaled measures for the change in $\hat{\beta}$ use the naive variance estimate which is a consistent estimate of the true variance of $\hat{\beta}$ only if the working correlation structure is correctly specified. Alternatively, scaling may be done by the robust variance estimate which is consistent even under misspecification of $R(\alpha)$ (Liang and Zeger, 1986). We prefer the naive variance estimate. The robust variance estimate, with squared residuals in the middle of its sandwich formula, is inflated by large residuals; so its use for standardization would mask observations with large residuals.

3.3.2. Assessing the influence of case-deletion on the fitted values

Diagnostics analagous to Cook (1977, 1979) can be obtained to measure the influence of observations on estimated values of the linear predictor, and hence the fitted values. These diagnostics measure the influence of the deleted subset m on the overall fit. The change in the overall fit is given by either $X\Delta\hat{\beta}_m$, or $X_{[m]}\Delta\hat{\beta}_m$, where $\Delta\hat{\beta}_m = \hat{\beta} - \hat{\beta}_{[m]}$, depending upon whether one's view is that of deleting or adding a subset of observations. A class of norms which are location and scale invariant is given by $\mathrm{D}_m(\mathrm{M};\mathrm{c}) = (\Delta\hat{\beta}_m)'\mathrm{M}(\Delta\hat{\beta}_m)/\mathrm{c}$. First, norms with $\mathrm{c} = p\phi$ and $\mathrm{M} = X'WX$ are considered.

THEOREM 2. *A measure of standardized influence of a subset of observations, denoted by m, on the linear predictor is given by*:

$$\begin{aligned} \mathrm{D_m}(X'WX; p\hat{\phi}) &= (\hat{\beta} - \hat{\beta}_{[m]})'(X'WX)(\hat{\beta} - \hat{\beta}_{[m]})/(p\hat{\phi}) \\ &= \tilde{E}_m'(\mathrm{W}_m^{-1} - \tilde{Q}_m)^{-1}\,\tilde{Q}_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{E}_m/(p\hat{\phi}). \end{aligned} \tag{3.8}$$

As in section 3.1, we consider cluster-deletion and observation-deletion.

COROLLARY 2.1 *The effect of the i-th cluster on the overall fit is*

$$DCLS_i = E_i'(W_i^{-1} - Q_i)^{-1}Q_i(W_i^{-1} - Q_i)^{-1}E_i/(p\hat{\phi}). \tag{3.9}$$

COROLLARY 2.2 *The effect of the t-th observation in the i-th cluster on the overall fit is*

$$DOBS_{it} = \frac{\tilde{E}_{it}^2\,\tilde{Q}_{it}}{p\hat{\phi}(W_{it}^{-1} - \tilde{Q}_{it})^2}.$$

The results in THEOREM 2, COROLLARY 2.1 and COROLLARY 2.2 follow directly from THEOREM 1, COROLLARY 1.1 and COROLLARY 1.2.

The Cook statistic for generalized linear models, (2.30), is like (3.8) but has a diagonal W. Under the identity link, but allowing dependence, COROLLARY 2.2 gives the Cook type measure, (2.37), of Christensen, Pearson and Johnson (1992); they refer to $\tilde{Q}_{it}W_{it}$ as the generalized leverage. For the multiple linear regression model, (3.8) reduces to the multiple case-deletion Cook's distance, (2.8). For a single observation, this is equivalent to the original Cook's distance (Cook, 1977; Cook 1979). A value of about 1.0 is generally considered large (Kleinbaum et al. (1988), p. 201). In section 4, we consider the presence of correlation in interpreting (3.9) through an illustrative example.

An alternative measure to that of THEOREM 2, can be achieved by scaling with the subset m deleted from the covariance matrix by letting $\mathrm{M} = X'_{[m]}W_{[m]}X_{[m]}$ or $\mathrm{M} = X'_{[m]}\mathrm{V}^{-1}_{[m]}X_{[m]}$. In general, these two choices are different because $V^{-1}_{[m]} = W_{[m]} - W_{[m]m}W_m^{-1}W_{m[m]}$. The exception is cluster deletion where $W_{[m]} = V^{-1}_{[m]}$, in which case applying COROLLARY 1.1 and the relation,

$$X'_{[i]}W_{[i]}X_{[i]} = X'WX - X'_iW_iX_i,$$

the studentized distance for the influence of the i-th cluster on the overall fit is

$$MCLS_i = E'_i(W_i^{-1} - Q_i)^{-1}H_iE_i/(p\hat{\phi}) \,. \tag{3.10}$$

Notice that $MCLS_i$ is expressed as the product of the cluster leverage and the squared residual scaled by var(E_i) as defined in section 2. The analagous measure for observation-deletion depends on the choice of M and is a bit more complicated algebraically. For multiple linear regression, $\mathrm{D}_m(X'_{[m]}X_{[m]};1)$ gives the multiple case-deletion diagnostic, $MDFFIT_m$, in (2.7) and $\mathrm{D}_i(X'_{[i]}X_{[i]};1)$ gives the single case-deletion, $DFFIT_i$, in (2.5).

In summary, diagnostics for assessing the influence of observations on the fitted values of the linear predictor may be scaled by the variance estimate of $\hat{\beta}$ based on all the observations as in (3.9) or on all but the subset of observations to be deleted as in (3.10). The former has the attraction that the comparison of distances between observations is meaningful because they refer to the same metric, as pointed out by Cook and Weisberg (1982) and Chatterjee and Hadi (1986) for linear regression. On the other hand, since the deleted case influences the estimate of the variance, its inclusion may decrease the magnitude of the diagnostic and, to some degree, may hide influence. Welsch (1986), prefers the studentized version, because of its "robustness". Cook (1986) points out, however, that the studentized diagnostic has a different interpretation than the standardized version. Thus the diagnostic given in (3.9) which is a generalization of Cook's distance (1977) has the interpretation of the influence of a subset of observations

on $\hat{\beta}$, whereas (3.10) which is a generalization of $DFITS$ of Belsley et al. (1980, p. 15) has the interpretation of the influence of a subset of observations on $\hat{\beta}$ and the variance estimate of $\hat{\beta}$ simultaneously. For this reason, the question, "influence on what" should be the determining factor in choice of $DCLS$ or $MCLS$.

3.4 Examples from Medicine

3.4.1 Medical Practice data

Data from the North Carolina Early Cancer Detection Program at the Lineberger Comprehensive Cancer Center are used to illustrate the use of the diagnostics presented in section 3. Chart review data were collected from a random sample of $N = 3889$ medical charts in $K = 57$ medical practices. A practice constitutes a cluster. The number of charts per practice ranges from 19 to 197. The response $Y_{it} = 1$ if the t-th chart in the i-th practice indicates that the patient made at least one "health maintainence visit" during the years 1990 and 1991, and $Y_{it} = 0$ otherwise. An exchangeable correlation stucture is assumed; $\text{corr}(Y_{it}, Y_{it'}) = \rho,\ i = 1, \ldots, K,\ t \neq t' = 1, \ldots, n_i$. We specify a logit link in (3.1) and consider the following covariates:

INTERCEPT	
SPECLTY	Doctor's specialty: 0 if family practice or general practitioner 1 if internal medicine,
MDAGE	(Doctor's age in years − 45)/10,
M65	Doctor's gender: 0 if male, 1 if female,
PATAGE	(Patient's age in years − 65)/10,
NOINSUR	Health insurance: 0 if insured, 1 if not insured,
NBRMDS	Number of doctors in practice − 1,
M3	(Number of patients over 50 years old seen per day − 15)/10,
M63	Doctor's flu vaccination: 0 if in the last two years, 1 if 3 to 5 years ago, 2 if never,

MALEPAT Patient's gender: 0 if female, 1 if male,

BLACKPAT Patient's race: 0 if white, 1 if black.

Column (a) of Table 3.1 shows parameter and standard error estimates. Because of large cluster sizes no single observation had a large influence, so we present cluster-deletion diagnostics only. Exact cluster-deletion diagnostics were obtained by iterating the fitting algorithm, including estimation of ρ, to convergence.

Figures 3.1, 3.2 and 3.3 show plots of one-step approximations versus their exact counterparts. Figures 3.1 and 3.2 show $DBETAC$ for M3 and SPECLTY, respectively, while Figure 3.3 shows the Cook's distance, or $DCLS$. First, the plots show that the one-step approximations given by (3.5) are in good agreement with the exact diagnostics. This was the case for the other coefficients, and with respect to (3.6), as well. Second, cluster 5 has a large influence on the slope for M3; and clusters 19 and 29 have a large influence on the slope for SPECLTY. Figure 3.3 shows that clusters 5, 15, 19 and 29 had the largest influence on the fitted values; the log scale is used for clarity.

Figure 3.4 shows a plot of $DCLS$ versus cluster size. Clearly, clusters 15 and 29 have a large influence relative to their size. Figure 5 shows that cluster 29 had less leverage than clusters 15 and 19 but about the same influence, indicating that it had large residuals, that is, it was poorly fit. Figure 3.6 shows the relationship of leverage to cluster size; cluster 5 had the largest leverage and cluster 48, although not influential with respect to (3.5), had a large leverage relative to its size, which was the smallest of all clusters.

Estimates obtained after deleting cluster 5 are shown in column (b) of Table 3.1. The estimates for M3 changed by 2.94 standard errors, while for M65 it changed by 1.18 standard errors. The one-step approximations, (3.7), shown in Table 3.2, are 3.21 and 1.15, respectively. Table 3.2 also shows important summary information for the most

influential clusters; only clusters 5, 19, 15 and 29 have at least one value of (3.7) greater than 1.0 in absolute terms.

3.4.2 Subset of Medical Practice data

In order to study influence when cluster sizes are small, we consider a data set which is a randomly drawn subset of the Medical Practice data set described above. The data was selected such that 19 or one third of the clusters have $n_i = 2$, one third have $n_i = 3$, and one third have $n_i = 4$. After cluster sizes were assigned we randomly selected the observations or charts from all those available for each practice. What resulted was a data set with N = 171 charts in K = 57 practices. Clusters of size 1 were not included in order to facilitate study of the influence of observations, as opposed to clusters. A model similar to the one in section 4.3.1 will be considered.

Column (a) of Table 3.3 reports the parameter and naive standard error estimates. Figure 3.7 plots the Cook's distance diagnostic, $DOBS_{it}$ versus its exact counterpart. Again, the plot is on the log scale for clarity. The plot shows that $DOBS_{it}$ is a good approximation to the exact quantity, and that no single observation stands out as having very large influence on the fit. The largest Cooks distance, in fact, was 0.0481 which corresponded to observation with i.d. number 4. The influence diagnostics of the four most influential observations are presented in Table 4.4. Figure 3.8 plots Cooks distance versus and observation leverage. It shows that observation 4 had the largest leverage, h_{it}, and the next three most influential observations with i.d.'s 17, 166, and 164, had large Pearson residuals. Column (b) of Table 3.3 reports the GEE parameter estimater and naive standard error estimates for the data without these four observations.

3.5 Discussion

In this chapter, one-step deletion diagnostics were introduced which identify influential data in the Generalized Estimating Equations procedure. The diagnostics are

good approximations of their exact counterparts, and their computation is fast, so that they may be routinely used in data analysis. In particular, for the example in section 3.4, (3.5) and (3.6) were approximately twenty-five times faster to compute than the exact quantities. Typically, diagnostics may identify faulty data which are subsequently removed from the analysis, otherwise the influence is tolerated. For the medical practice data in section 3.4.1, an intermediate solution of downweighting influential data using a robust regression may be better than removing 5% of the data corresponding to cluster 5. Such a procedure like that of Pregibon (1982), but for correlated outcomes, will be pursued in chapter 4.

Table 3.1 *GEE Parameter estimates and naive standard errors with* (a) *and without* (b) *cluster 5*

	(a)			(b)		
	$\hat{\beta}_j$	se	Z	$\hat{\beta}_j$	se	Z
intercept	-0.044	0.164	-0.269	0.045	0.174	0.258
speclty	-0.475	0.152	-3.133	-0.597	0.163	-3.670
mdage	-0.311	0.055	-5.631	-0.281	0.061	-4.606
m65	0.533	0.154	3.451	0.350	0.158	2.213
patage	-0.104	0.031	-3.383	-0.103	0.031	-3.307
noinsur	-0.447	0.102	-4.386	-0.461	0.103	-4.475
nbrmds	0.025	0.059	0.431	0.037	0.066	0.562
m3	0.480	0.090	5.316	0.214	0.111	1.932
m63	0.001	0.061	0.022	-0.034	0.065	-0.530
malepat	-0.400	0.067	-5.959	-0.424	0.068	-6.245
blackpat	-0.441	0.126	-3.490	-0.369	0.133	-2.778

(a) based on all data: $N = 3889$ patients; $K = 57$ medical practices; $\hat{\rho} = 0.154$.

(b) cluster 5 deleted: $N = 3698$ patients; $K = 56$ medical practices; $\hat{\rho} = 0.171$.

Table 3.2 *Cluster size, leverage and influence diagnostics for selected clusters and covariates*

				$DBETACS_{ij}$			
cluster	n_i	tr(H_i)	$DCLS_i$	SPECLTY	M65	M3	M63
5	191	1.26	1.25	0.49	1.15	3.21	0.74
19	197	0.87	0.36	-1.85	-0.76	0.29	0.27
15	158	0.58	0.33	0.33	-0.37	-0.78	1.09
29	89	0.32	0.27	-1.20	1.07	0.47	0.23
48	19	0.26	0.02	-0.01	-0.16	-0.20	-0.19

$DCLS$ is given by (11) in section 3 · 2; $DBETACS$ is given by (8) in section 3 · 1.

Table 3.3 *GEE Parameter estimates and naive standard errors with* (a) *and without* (b) *observations 4, 17, 164 and 166 for the subset of Medical Practice Data*

	(a)			(b)		
	$\hat{\beta}_j$	se	Z	$\hat{\beta}_j$	se	Z
intercept	0.594	0.360	1.650	0.663	0.392	1.691
speclty	0.266	0.387	0.687	0.415	0.435	0.954
mdage	-0.326	0.232	-1.405	-0.399	0.257	-1.553
m65	-0.044	0.496	-0.089	0.169	0.550	0.307
patage	-0.068	0.176	-0.386	-0.043	0.184	-0.234
noinsur	0.672	0.577	1.165	1.424	0.662	2.151
nbrmds	-0.263	0.114	-2.307	-0.331	0.133	-2.489
m3	0.004	0.370	0.011	-0.143	0.418	-0.342
m63	-0.290	0.277	-1.047	-0.419	0.310	-1.266
malepat	-1.342	0.366	-3.667	-1.658	0.383	-4.329
blackpat	-0.602	0.487	-1.236	-0.955	0.537	-1.778

(a) based on all data: $N = 171$ patients; $\hat{\rho} = 0.095$.

(b) observations 4, 17, 164, and 166 deleted: $N = 167$ patients; $\hat{\rho} = 0.179$.

Figure 3.1 1-Step versus exact change in parameter coefficient for M3

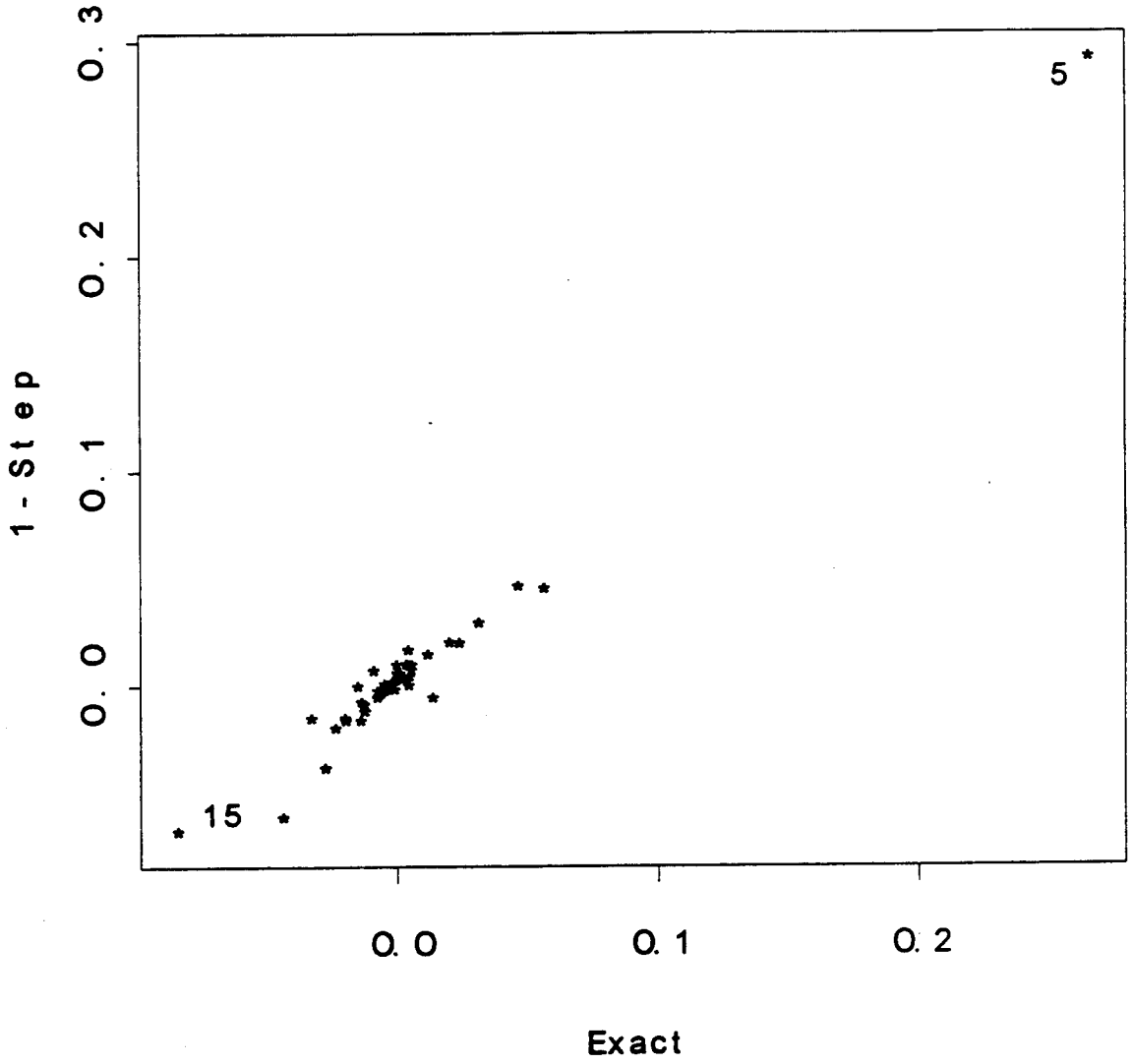


Figure 3.2 1-Step versus exact change in parameter coefficient for SPECLTY

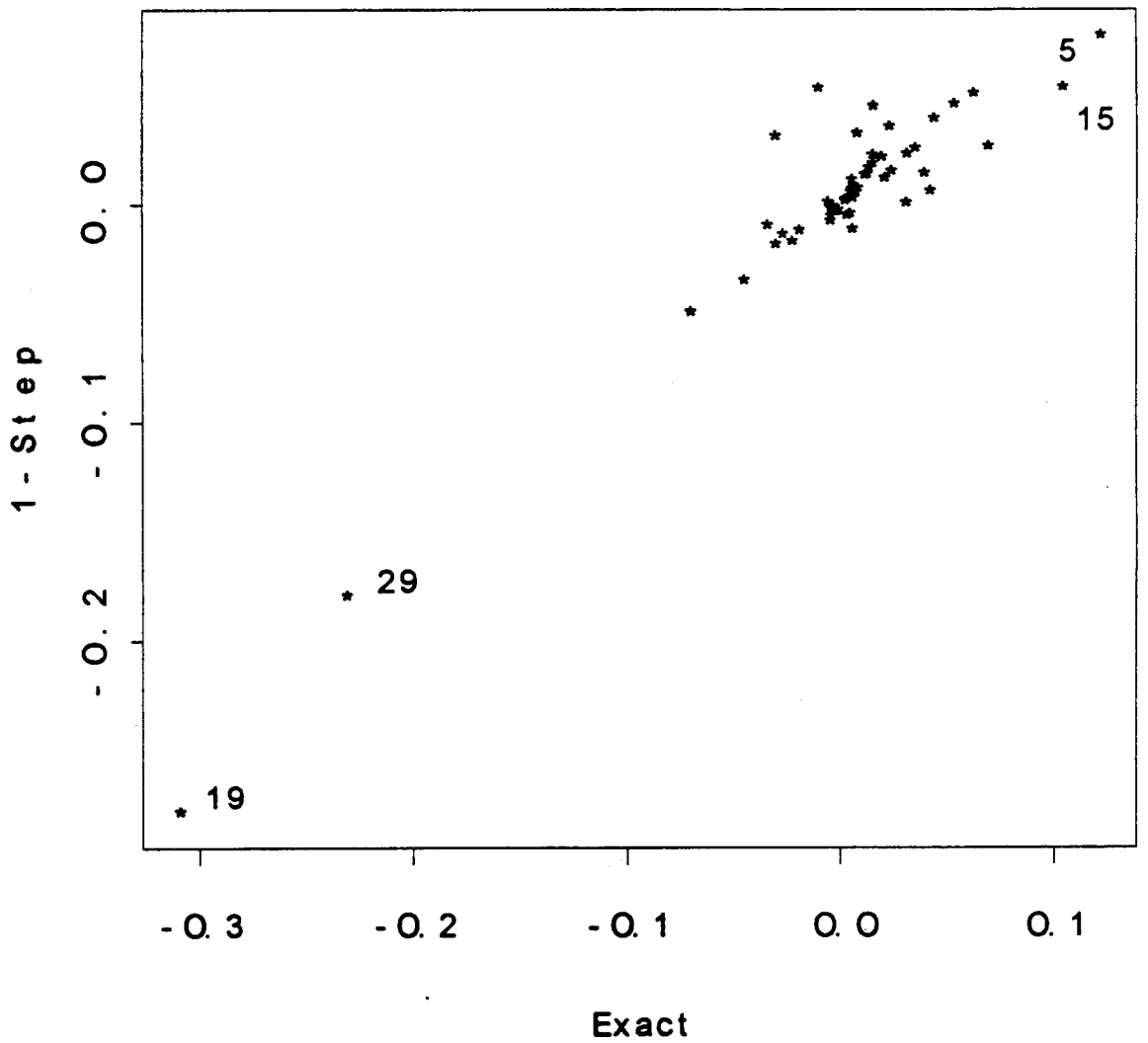


Figure 3.3 1-Step versus exact value of log of Cook's Distance for clusters

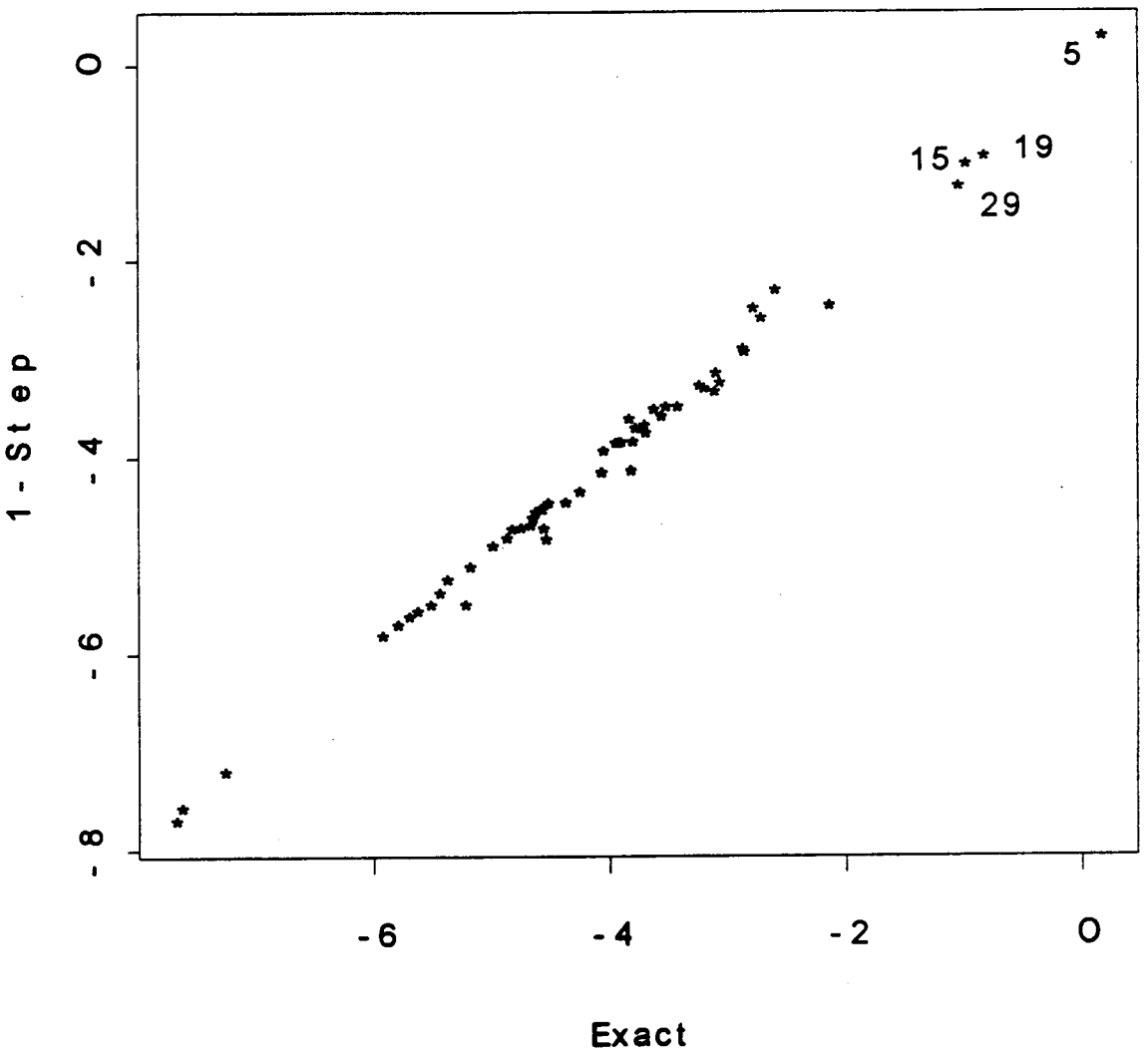


Figure 3.4 Cook's Distance versus cluster size

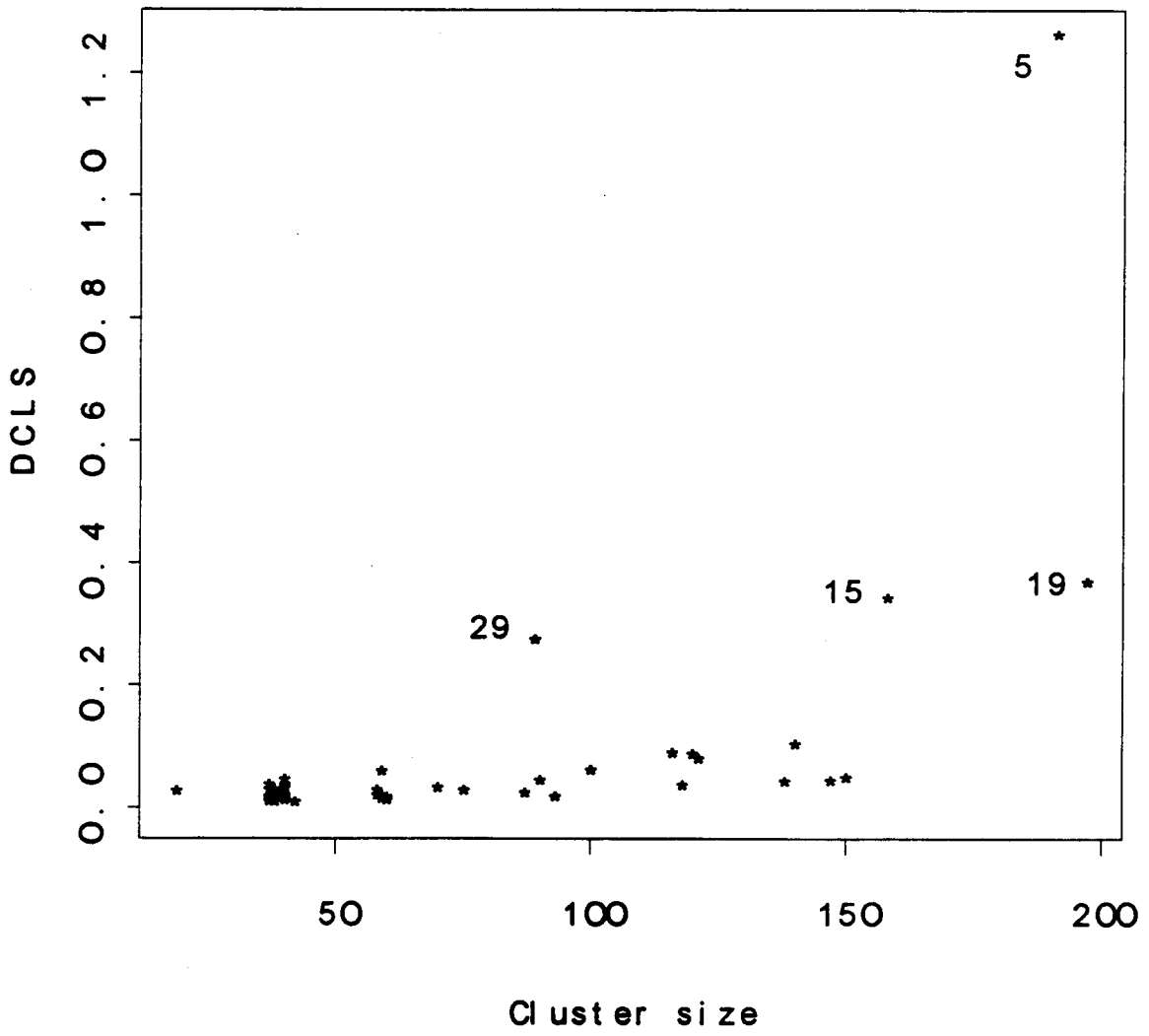


Figure 3.5 Cook's Distance versus cluster leverage

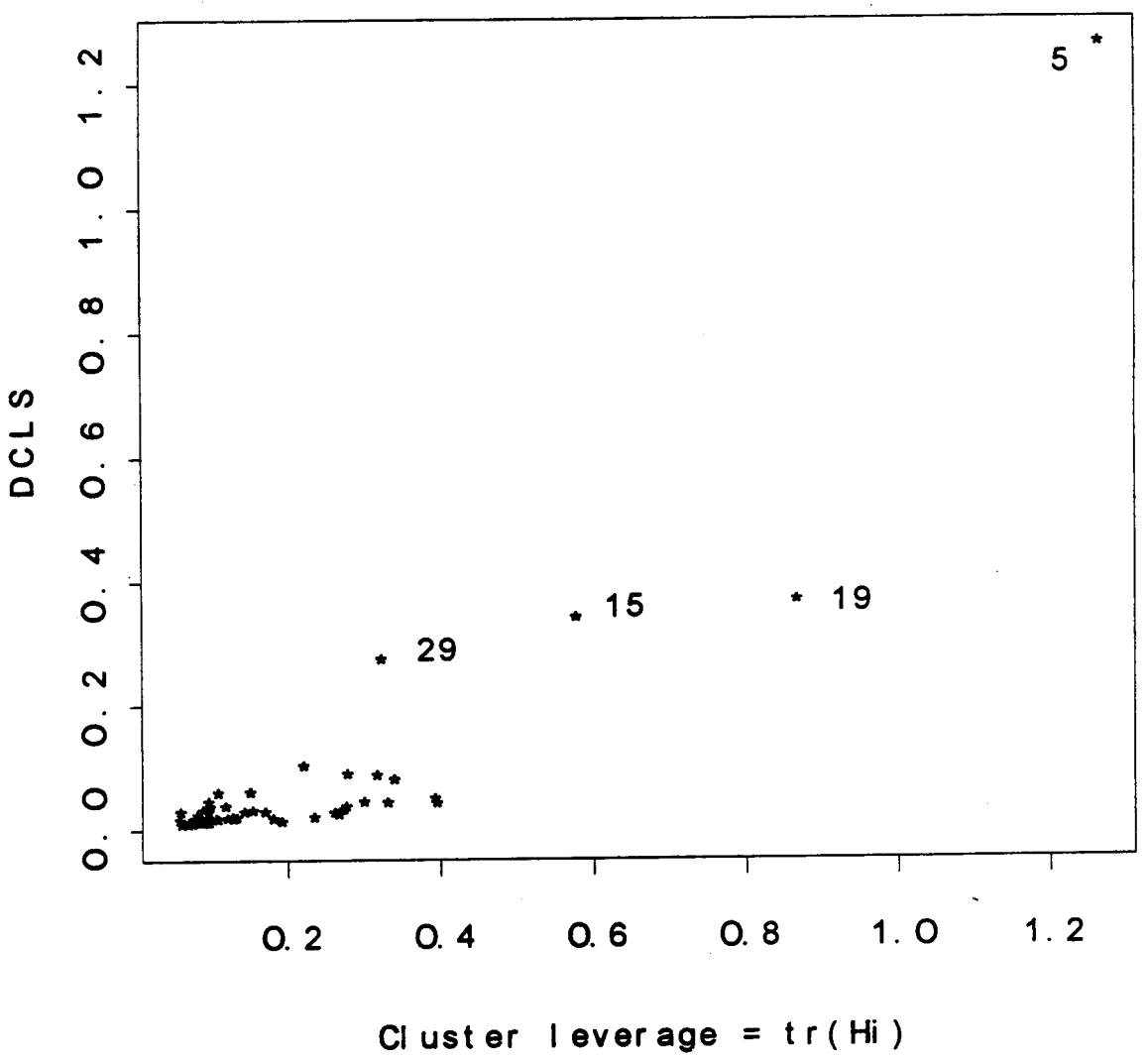


Figure 3.6 Cluster leverage versus cluster size

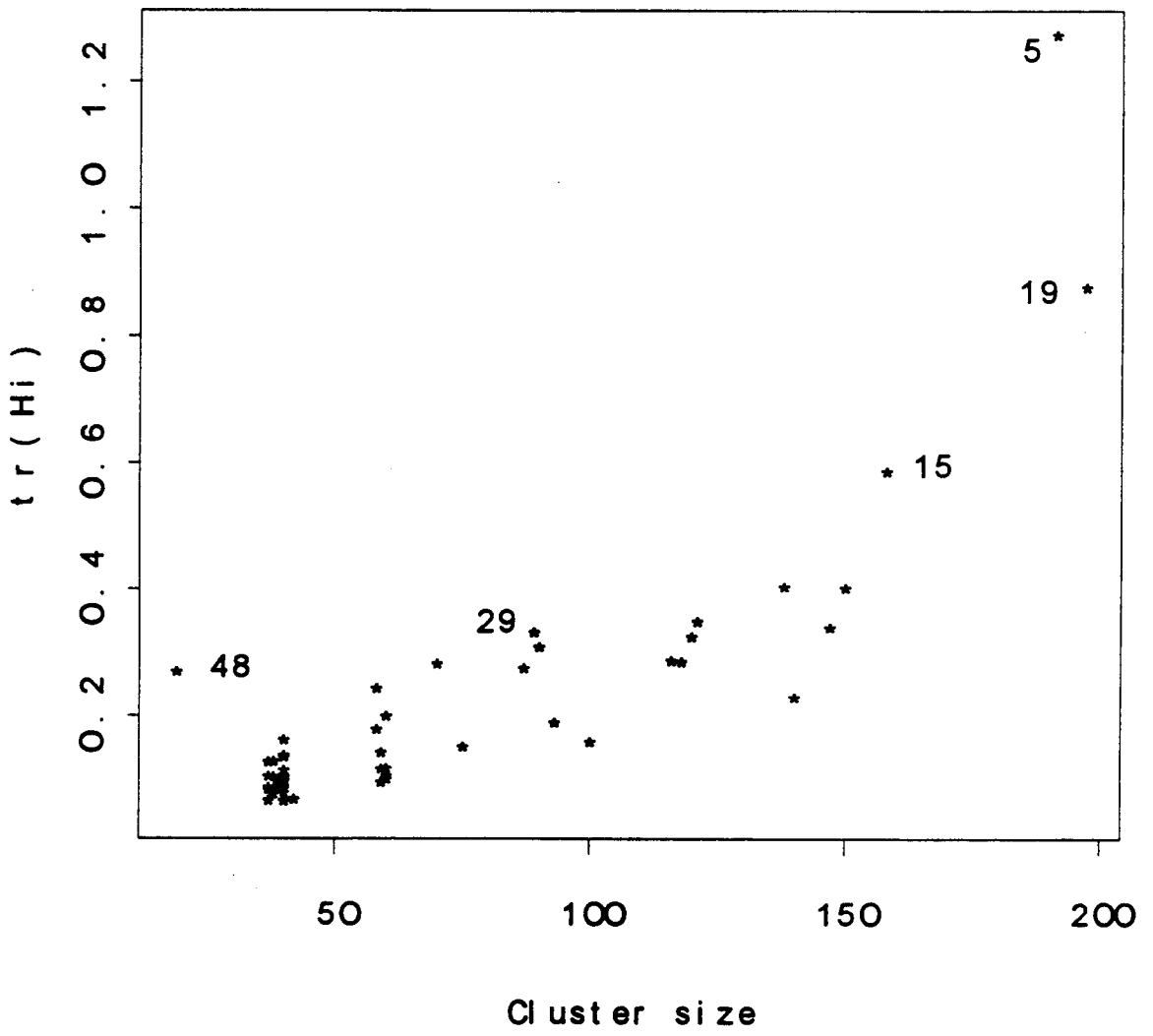


Figure 3.7 1-Step versus exact value of log of Cook's Distance for observations

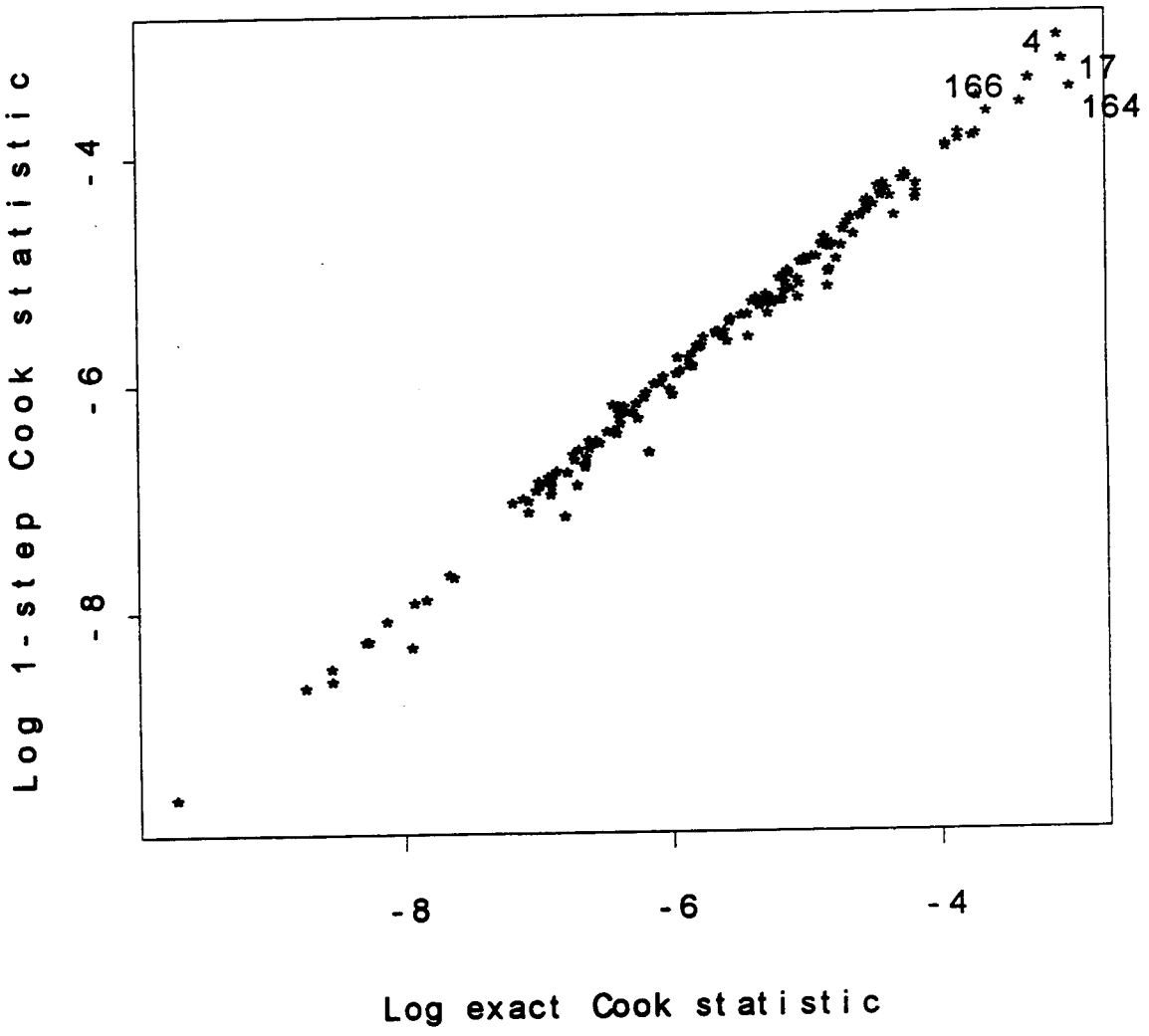
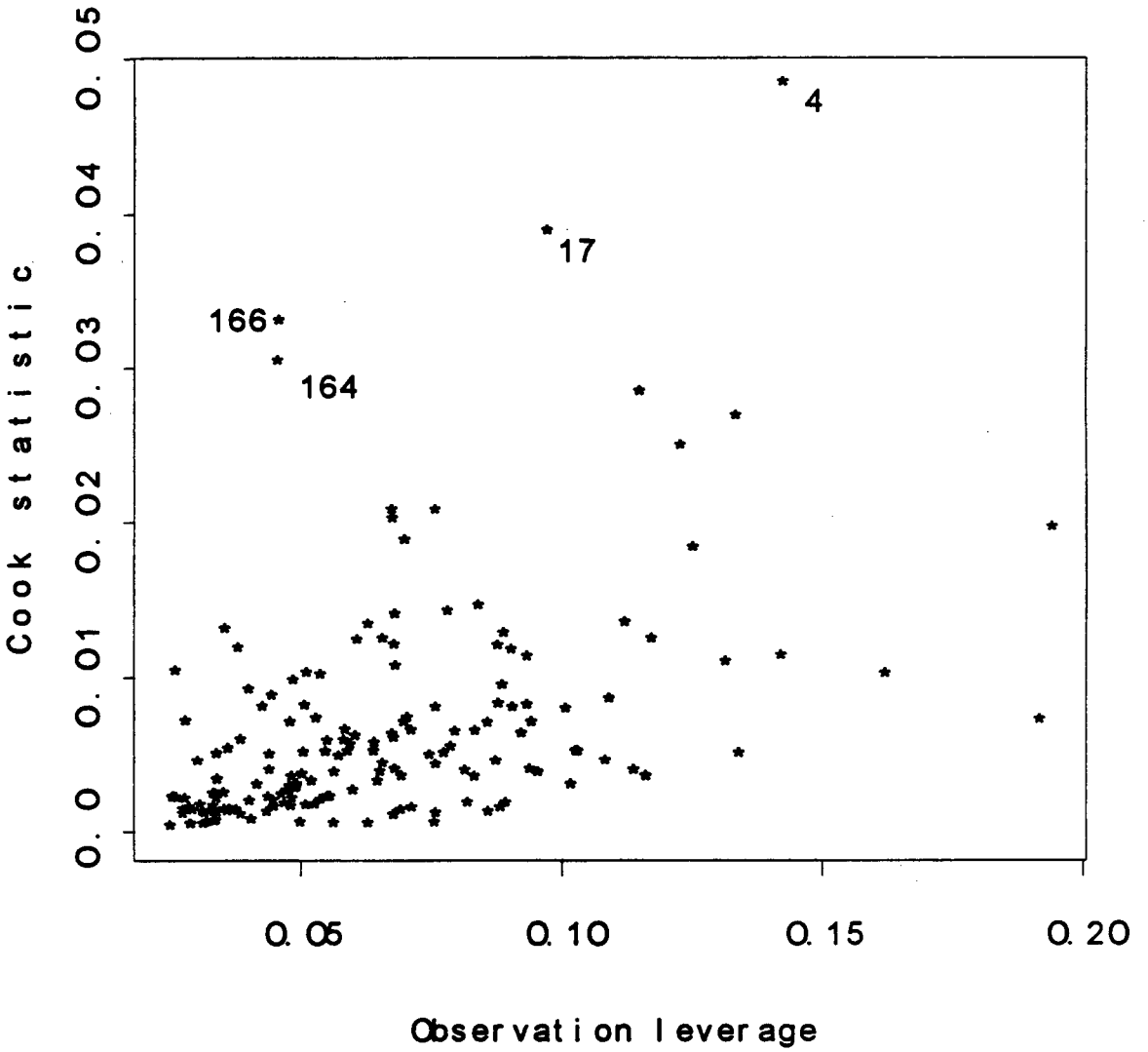


Figure 3.8 Cook's Distance versus observation leverage



CHAPTER IV

RESISTANT GENERALIZED ESTIMATING EQUATIONS

4.1 Introduction

In Chapter 3, examples were given where the parameter estimates from an analysis based on the GEE procedure were highly influenced by a small subset of the data. A resistant fit, on the other hand, is one which is not sensitive to large changes in a few observations (Pregibon, 1982). This chapter introduces Resistant Generalized Estimating Equations (REGEE) which give regression parameter estimates that are resistant to the influence of observations or clusters. They are a modification of those of Liang and Zeger (1986) and reduce to theirs when there are no unusual observations such that all observations receive equal weight. Our robust regression approach is a generalization of Carroll and Pederson (1993) who provide robust estimates in the logistic regression model that are of the Mallows class. Estimates of the Mallows class are obtained by downweighting observations with large leverage values. Alternatively, Schweppe estimates are obtained by downweighting observations according to their residuals as in Pregibon (1982) and Künsch et al. (1989). Within both classes, we consider two approaches which we call observation-downweighting and cluster-downweighting. The former downweights individual observations separately, whereas the latter method assigns equal weight to all observations in a cluster based on some aggregate measure of the influence of the entire cluster. Singer and Sen (1985) and Huggins (1993) have proposed robust multivariate methods for continuous responses. Our methods have wider applicability in that they apply to the same class of models

considered by GEE. We consider the application of REGEE to correlated binary data and discuss their potential and limitations to other types of outcomes.

4.2 Resistant approach to modeling correlated outcomes

4.2.1 Resistant Generalized Estimating Equations

In order to provide robust estimation of the broad class of models considered by Liang and Zeger (1986), including the medical examples in section 2, we define Resistant Generalized Estimating Equations (REGEE) as

$$\sum_{i=1}^{K} D_i'(X_i,\beta)V_i^{-1}(\alpha,\beta)\Big[W_i(X_i,X,Y_i,\alpha,\beta)(Y_i-\mu_i(\beta))-c_i\Big] = 0 \qquad (4.1)$$

where $V_i = A_i R_i(\alpha) A_i$, and D_i, R_i and A_i are defined as in (3.2), and μ_i is parametrized as in (3.1). Note, however, that V_i and W_i are defined differently in this chapter than in chapter 3. The GEE equations given by (3.2) are a special case of (4.1) in which $W_i = \mathrm{I}$ and $c_i = 0$. In general, however, W_i is a diagonal matrix for the i-th cluster which contains weights, w_{it}, t = 1, ... , n_i, that correspond to the elements of the reponse vector, Y_i. Downweighting may be done based on the covariates only, the so-called Mallows class, or on the responses as well, the Schweppe class. For the Mallows classs, $c_i = 0$. For the Schweppe class c_i is determined so that (4.1) are unbiased estimating equations. The w_{it} are between 0 and 1 and are analagous to the ψ - functions in M-estimation in section 2.1.2 in that they determine the robustness and efficiency of $\widehat{\beta}$. Most observations will have a weight of or near 1, but those observations which are determined to have large influence on the estimation of β will receive a smaller weight. The following Theorem gives asymptotic results for (4.1) under the model in (3.1).

Theorem 3 : *Define* $\psi_i = W_i(Y_i - \mu_i)$ *and* $c_i = \mathrm{E}[\psi_i]$. *Assume that* :

(*i*) $\widehat{\alpha}$ *is* $k^{1/2}$ *– consistent given* β *and* ϕ;

(*ii*) $\widehat{\phi}$ *is* $k^{1/2}$ *– consistent given* β;

(iii) $Var(\psi_i) < \infty$;

(iv) ψ_i *is absolutely continuous in* β *such that the derivative with respect to* μ_i, *denoted* $\dot{\psi}_i$, *exists, for almost all y, and* $\mathrm{E}\| \dot{\psi}_i \| < \infty$;

Under additional regularity conditions, $k^{1/2}(\widehat{\beta}_S - \beta)$ *is asymptotically Gaussian with zero mean and covariance matrix* V_S *given by*

$$\lim_{k \to \infty} k(\sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i)^{-1} \{\sum_{i=1}^{k} D_i' V_i^{-1} \mathrm{Var}(\psi_i) V_i^{-1} D_i\} (\sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i)^{-T} \qquad (4.2)$$

where $\Gamma_i = \mathrm{E}\dot{\psi}_{ki} - \dot{c}_i$, $\dot{\psi}_{ki} = \frac{\partial}{\partial \mu_i} \psi_i(\mu_i)$ *and* $\dot{c}_i = \frac{\partial}{\partial \mu_i} c_i$.

The additional regularity conditions are that derivatives, $\partial \widehat{\alpha}(\beta, \phi)/\partial \phi$, $\partial \widehat{\alpha}\{\beta, \widehat{\phi}(\beta)\}/\partial \beta$, and $\partial \widehat{\phi}(\beta)/\partial \beta$, are bounded in variation. The proof for Schweppe observation downweighting REGEE is in Appendix 2. A sketch of the proof follows. The proof for the special case of the Mallows class is found in Appendix 3, although the general proof applies as well.

Sketch of Proof: The sketch here considers α and ϕ are known; REGEE for the Schweppe class may be written as

$$\mathrm{U}(\beta) = \sum_{i=1}^{K} \mathrm{U}_i(\beta) = \sum_{i=1}^{K} D_i'(\beta) V_i^{-1}(\beta) \left[\psi_{ki} - c_i\right]$$

where $\psi_{ki} = W_{ki}(Y_i, \beta)(Y_i - \mu_i(\beta))$ and $c_i = \mathrm{E}(\psi_{ki})$.

A Taylor expansion gives

$$\mathrm{U}(\widehat{\beta}) = 0 = \mathrm{U}(\beta) + \partial \mathrm{U}(\beta)/\partial \beta \Big|_{\beta = \widetilde{\beta}} (\beta - \widehat{\beta})$$

for some $\widetilde{\beta} = \lambda \beta + (1 - \lambda)\widehat{\beta}$, $\lambda \epsilon (0, 1)$. It follows that,

$$k^{1/2}(\widehat{\beta}-\beta)=\left[\frac{\partial U(\widetilde{\beta})/\partial\beta}{k}\right]^{-1}\left(U(\beta)/k^{1/2}\right).$$

By the Central Limit Theorem for triangular arrays (Theorem 3.3.5 in Sen and Singer; 1993) in conjunction with the Cramer-Wold device, and regularity conditions, and assumptions (*iii*) and (*iv*) above,

$$U(\beta)/k^{1/2}\sim MVN\left(0,\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}D_i'V_i^{-1}\mathrm{Var}(\psi_i)V_i^{-1}D_i\right).$$

Note that assumption (*iii*) follows from: (1) the fact that the weight matrix W_{ki} is bounded between 0 and 1, and (2) the usual GEE assumptions of the finiteness of the second moment of Y_i. Next, it can be shown that,

$$E[\partial U(\beta)/\partial\beta]=\sum_{i=1}^{k}D_i'V_i^{-1}\Gamma_iD_i$$

To see this note that $\partial U_i(\beta)/\partial\beta=\frac{\partial}{\partial\beta}\{D_i'V_i^{-1}\}\left[\psi_{ki}-c_i\right]+\{D_i'V_i^{-1}\}\frac{\partial}{\partial\beta}\left[\psi_{ki}-c_i\right]$.

The first part has expectation zero, and in the second part, under condition (*iv*),

$$\frac{\partial}{\partial\beta}\left[\psi_{ki}-c_i\right]=\frac{\partial}{\partial\mu}\left[\psi_{ki}-c_i\right]\frac{\partial\mu_i}{\partial\beta}=\left[\dot{\psi}_{ki}-\dot{c}_i\right]D_i.$$

Then by the Markow Weak Law of Large Numbers,

$$\frac{1}{k}\partial U(\beta)/\partial\beta\xrightarrow{p}\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}D_i'V_i^{-1}\Gamma_iD_i.$$

Since, under certain regularity conditions (see Appendix 2),

$$\frac{1}{k}\left|\left|\,\partial U(\widetilde{\beta})/\partial\beta-\partial U(\beta)/\partial\beta\,\right|\right|\xrightarrow{p}0,$$

it follows that,

$$\frac{1}{k}\partial U(\widetilde{\beta})/\partial\beta\xrightarrow{p}\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}D_i'V_i^{-1}\Gamma_iD_i.$$

Lastly, by Slutsky's Theorem, $k^{1/2}(\widehat{\beta} - \beta) \sim MVN(0, V_S)$.□

Theorem 3 implies that use of (4.1) will result in some loss of efficiency under model (3.1). However, chapter 5 will show that robustness is gained because the asymptotic results will apply to a broader class of models than (3.1), including models that allow for contamination.

Estimation of the regression parameter and variance estimate.

Estimation of β is done with iteratively reweighted least squares by regressing the working response vector $Z = X\widehat{\beta} + D^*(\psi - c)$ on X with the original weight matrix W^* from GEE. The variance of $\widehat{\beta}_S$ is estimated by

$$(\sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i)^{-1} \{\sum_{i=1}^{k} D_i' V_i^{-1} (\psi_i - c_i)(\psi_i - c_i)' V_i^{-1} D_i\} (\sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i)^{-T}, \quad (4.3)$$

The "robust" variance estimator of GEE, described in section 3.2, is obtained by setting $W_i = \mathrm{I}$ in (4.3).

Estimates of the Mallows observation-downweighting class are obtained by specifying weights $w_{it} = w_{it}(h_{it})$ which are updated at each iteration. An alternative way of measuring leverage in which weights are calculated only once is given by Carroll and Pederson (1993) in the logistic regression case. In the Schweppe observation-downweighting class, observations are downweighted according to their residual, and possibly, their leverage as well by specifying $w_{it} = w_{it}(x_{it}, X, y_{it}, \alpha, \beta)$. In this case, $c_i = \mathrm{E}[\psi_i]$ is determined by the marginal distributions of Y_{it}, so that (4.1) yields consistent estimates of β under Theorem 3 above. In particular, we consider weights $w_{it}(r_{it})$ that are a function of the Pearson residual, (2.27), with a robust estimate of scale, $\hat{\phi}$, as described in the next section.

In both Mallows and Schweppe classes, we may want to downweight clusters instead of observations, by assigning all the observations in a cluster equal weight. Mallows cluster-downweighting may be achieved by assigning $w_{it} = w(tr(H_i))$. For Schweppe cluster-downweighting, one possibility is to summarize the lack of fit of the observations in the cluster by downweighting clusters as a function of $r_i' r_i / (\widehat{\phi} n_i)$ where $r_i = (r_{i1}, \ldots, r_{in_i})'$.

Estimation of nuisance parameters.

In robust regression, the estimation of nuisance parameters must also be robust. In REGEE, residuals are inflated due to downweighting in the robust estimation of β. The method of moments is applied to $r_{it}^* = (\psi_{it} - c_{it})/\mathrm{v}_{it}^{1/2}$, where $c_{it} = \mathrm{E}[\psi_{it}]$ instead of r_{it} to obtain robust estimates of ϕ and α. The same weights that were used in (4.1) may define r_{it}^*. If the covariance of ψ_i is such that

$$\mathrm{Var}(\psi_i) = B_i \mathrm{Var}(Y_i) B_i \quad \text{where} \quad B_i = Diag\{b_{it}\}, \tag{4.4}$$

the robust estimators of scale and exchangeable correlation are

$$\hat{\phi} = \sum_{i=1}^{K}\sum_{t=1}^{n_i} r_{it}^{*2} \Big/ \left\{ \sum_{i=1}^{K}\sum_{t=1}^{n_i} b_{it}^2 - p \right\} \tag{4.5}$$

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \sum_{i=1}^{K}\sum_{t>t'} r_{it}^* r_{it'}^* \Big/ \left\{ \sum_{i=1}^{K}\sum_{t>t'} b_{it} b_{it'} - p \right\}. \tag{4.6}$$

If all the weights are 1, then (4.5) and (4.6) reduce to the GEE nuisance parameter estimates of Liang and Zeger (1986); in these estimates as in ours, the correction factor p adjusts for the estimation of p regression coefficients.

Mallows versus Schweppe and Observation versus Cluster downweighting

The following remarks can be made about the different classes of REGEE:

1. The Mallows class, where the weights are nonrandom, does not require additional assumptions about the univariate marginal distributions beyond (3.1) and applies to any

situation that GEE might be used, including modelling binary data, count data (Poisson), and continuous responses (normal or gamma variance functions). The Mallows class, either observation-downweighting or cluster-downweighting, is a special case of the Theorem in which $\Gamma_i = W_i$ and $\mathrm{Var}(\psi_i) = W_i \mathrm{Var}(Y_i) W_i$. As a result, (4.4), (4.5) and (4.6) apply with $b_{it} = w_{it}$. In (4.3), note that $(\psi_i - c_i) = W_i(\mathrm{Y}_i - \hat{\mu}_i)$.

2. The Schweppe observation-downweighting class requires full specification of the marginal univariate distributions for estimation of β and ϕ. This generally applies to independent responses too (Morgenthaler, 1992); an exception is the location and scale families of distributions such as the normal distribution. Although the full set of bivariate distributions are required for estimation of α, an independence working correlation structure can be used as in GEE. If the correlation is misspecified, $\hat{\beta}$ is consistent. Otherwise, if (4.4) holds, knowledge of complete bivariate distributions can be used for estimation of α by (4.6).

3. In the Schweppe cluster downweighting class, the weights depend on the full response vector for the cluster and, thus, the full multivariate distribution is required in order to calculate the debiasing factor, c_i, in (4.1). Unfortunately, multivariate generalizations of the Poisson and Gamma distributions and many other distributions in the exponential family present two limitations. First, computing c_i is very complicated and generally does not have a closed form expression. Second, additional parameters beyond β, α, and ϕ are required for calculation of c_i. Notable exceptions are the bivariate binary distribution and the multivariate normal distribution which are completely specified by β, α, and ϕ. However, even in the latter case, the calculation of c_i may be formidable.

4.2.2 REGEE for Correlated Binary Responses

The Schweppe observation-downweighting theory is easily applied to correlated binary outcomes because the marginal distributions and the bivariate distributions only depend on α and β. For a cluster, define $\pi_{jk} = Pr(Y_1 = j, Y_2 = k)$. The bivariate distribution is a multinomial distribution with cell probabilities (π_{11},π_{10},π_{01},π_{00}) which are completely determined by the marginal means $\mu_1 = Pr(Y_1 = 1)$, and $\mu_2 = Pr(Y_2 = 1)$ and the correlation between Y_1 and Y_2 denoted by ρ. Specifically, $\pi_{11} = \mu_1\mu_2 + \rho\, \mathrm{v}_1^{1/2}\mathrm{v}_2^{1/2}$, $\pi_{10} = \mu_1 - \pi_{11}$, $\pi_{01} = \mu_2 - \pi_{11}$, and $\pi_{00} = 1 - \mu_1 - \mu_2 + \pi_{11}$. The variance function is defined as $\mathrm{v}_t = \mathrm{v}(\mu_t) = \mu_t(1 - \mu_t)$ and $\phi = 1$.

Let the observation weight w_t is a function of the corresponding residual r_t. To solve the Schweppe observation-downweighting ReGEE, c_t is determined from the marginal Bernoulli distribution. It can be shown that $c_t = \mathrm{v}_t(w_t^{(1)} - w_t^{(0)})$, where $w_t^{(j)}$ is the weight for the t-th observation in the cluster evaluated at $\mathrm{y_t} = \mathrm{j}$. Now, because each w_t is a function of its corresponding residual and not the entire residual vector for the cluster, it follows that $\dot{\psi}$, and thus Γ are diagonal matrices. It can be shown that $\Gamma = Diag\{-b_t\}$, $\mathrm{Var}(\psi_t) = \mathrm{v}_t\, \mathrm{b}_t^2$, and $\mathrm{Cov}(\psi_t, \psi_{t'}) = \rho_{tt'}\, \mathrm{v}_t^{1/2}\mathrm{v}_{t'}^{1/2}\mathrm{b}_t\mathrm{b}_{t'}$, where $b_t = (1 - \mu_t)w_t^{(1)} + \mu_t w_t^{(0)}$, and $\rho_{tt'}$ is the correlation between $\mathrm{Y_t}$ and $\mathrm{Y_{t'}}$. Thus (4.4) holds and the exchangeable correlation parameter is estimated by (4.6). The conditions of Theorem 3 are verified for this example in Appendix 4.

The application of Schweppe cluster-downweighting for bivariate binary data is described in Appendix 5. Beyond clusters of size two difficulties arise due to the complexity of the multivariate distribution. The quadratic exponential distribution has been suggested as an alternative (Prentice and Zhao, 1991). Aside from making simplifying assumptions about higher moments, it has the unattractive quality of not

being reproducible. In other words, marginal distributions do not belong to the same family as the multivariate distribution (Cox and Wermuth, 1994).

Finally, for binary data, our result generalizes existing theory for robust logistic regression. For a logit link applied to independent binary data, the Mallows class theory from Theorem 3 is equivalent to (2.34). For the Schweppe class, our approach is related to that of Künsch et al. (1989).

4.3 Examples from Medicine revisited

The weight function applied throughout this work is $w(v) = exp(-(v/a)^2)$ (Holland and Welsch (1977)). Values of a, the tuning constant, depend on the class of REGEE and on v. Choices of a are explained in the applications below. In general, the tuning constants applied here, always resulted in a mean weight above 0.90 and no more than 10% of the observations receiving weight less than 0.75. Tuning constants are not too large such that potentially influential data are not downweighted.

4.3.1 Medical Practice data

Mallows cluster-downweighting with $v = tr(H_i)$ and $a = 1.2$ is applied to the Medical Practice data described in section 3.4.1. This tuning constant gives a cluster with $v = 1.0$ a weight, $w = .5$. REGEE estimates and standard errors are given in column (c) of Table 4.1. The estimate of M3 is 0.333 which is between the estimates in columns (a) and (b). The change in M3 is primarily caused by the heavy downweighting ($w = 0.32$) of cluster 5. Table 4.2 shows the final weights corresponding to the four influential clusters. The three most influential clusters were downweighted because of their large leverage. Cluster 29, however, was not downweighted nearly as much because as pointed out in section 3.4.1, its leverage does not contribute that much to its influence. All clusters not shown in Table 4.2 had weights exceeding 0.89 and the median weight of the 57 clusters was 0.99. In summary, the Mallows cluster-

downweighting approach provides an attractive alternative to the non-resistant GEE estimates in column (a) of Table 4.1 and to the estimates in column (b) which are due to the removal of 5% of the data which is cluster 5.

4.3.2 Subset of Medical Practice data

Both the Mallows observation-downweighting and the Schweppe observation-downweighting REGEE were applied to the data described in section 3.4.2. Column (b) of Table 4.3 shows the Mallows estimates and standard errors when $v = h_{it}$ and $a = 0.19$; this tuning constant is equal to three times the mean of the observation leverages, i.e., $a = 3p/N$. Column (c) of Table 3 shows the Schweppe estimates and standard errors when $v = r_{it}$ and $a = 3.0$. For both approaches an observation with $v = a$ will receive a weight $w_{it} = 0.5$. Table 4 contains the final weights assigned to the four observations having the highest influence as indicated by $DOBS_{it}$. On the whole, the Mallows approach was not very effective. The most influential observation was downweighted ($w = 0.65$) although its weight was only the eighth smallest. The next three influential observations were hardly downweighted at all. In the Mallows approach, NBRMDS has Z-score of -1.90 which would not be significant at the .05 level, and would lead to a qualitatively different conclusion than GEE or the Schweppe approach where Z = -2.03. The Schweppe approach performed better. Although the most influential observation was only downweighted mildly ($w = 0.78$), the next three observations received the smallest weights among all observations. The Schweppe approach was more effective than the Mallows approach because influence in this data set is more related to residual than leverage. In summary, the influence of any single observation was moderate at best, the magnitude of the estimates across the three columns of Table 4.3 changed very little, and we conclude that GEE provides satisfactory estimates for the model.

4.4 Discussion

In this chapter, we have presented an alternative methodology which generalizes the GEE approach by downweighting influential clusters or observations. We have shown that individual observations or clusters may have a large impact on GEE parameter estimates. REGEE on the other hand is highly resistant to their influence. It is comparable to GEE in computation time because the fitting process is by iterative weighted least squares as in GEE. For discrete responses, the concepts of breakdown point and influence function of Hampel et al. (1986) are not applicable. First, even for logistic regression, all robust estimates in the literature have unknown breakdown point (Christmann, 1994). Second, the definition of influence function requires the full distribution, in our case the multivariate distribution, to be known. This is often not the case. In this paper, instead of examining robustness of REGEE in a mathematical sense, we have shown its robustness properties by illustrating its capability to downweight clusters or observations identified as influential by GEE regression diagnostics.

It is possible to apply REGEE in many ways not investigated here. Working correlation structures other than exchangeable would lead to correlation estimates different than (4.6). Even in the exchangeable case, a different, although asymptotically equivalent formula, can be obtained by the method of moments based on r^*_{it}/b_{it} instead of r^*_{it} in section 4.2. Finite sample properties of competing estimates of α in REGEE and GEE need further investigation. In this paper, we considered one weight function only, but others satisfying the Theorem may be used as well. In a similar framework, it may be possible to apply Huber's function (see Pregibon, 1982) and others that do not satisfy assumption (iv) in the Theorem. Finally, for the second example in section 4.3, it is possible to consider a Schweppe observation weight that is a function not only of residual, but leverage as well as in Pregibon (1982).

Special consideration is needed according to the class of REGEE being applied. Mallows REGEE applies to the broad class of models considered by GEE. The beauty,

and practicality, of GEE, and of Mallows ReGEE, is that only a working correlation structure along with the marginal means and variances (up to a scale parameter) are required in order to fit the model of interest. However, in order to gain robustness against isolated cases, stronger distributional assumptions are needed in the Schweppe class of ReGEE. For Schweppe observation-downweighting, the bias must be calculated from the marginal distributions. Multivariate distributional assumptions are required for the Schweppe cluster-downweighting REGEE. In summary, REGEE provides an attractive and feasible alternative to GEE because it automatically produces parameter estimates which are resistant to influential data.

Table 4.1 *Parameter estimates and standard errors for the Medical Practice data using (a) GEE, (b) GEE without cluster 5, and (c) Mallows cluster-downweighting REGEE.*

	(a)		(b)		(c)	
	$\hat{\beta}_j$	se	$\hat{\beta}_j$	se	$\hat{\beta}_j$	se
intercept	-0.044	0.213	0.045	0.211	-0.043	0.200
speclty	-0.475	0.295	-0.597	0.296	-0.418	0.299
mdage	-0.311	0.064	-0.281	0.089	-0.305	0.075
m65	0.533	0.286	0.350	0.258	0.458	0.252
patage	-0.104	0.034	-0.103	0.036	-0.105	0.036
noinsur	-0.447	0.118	-0.461	0.121	-0.458	0.122
nbrmds	0.025	0.053	0.037	0.060	0.035	0.058
m3	0.480	0.216	0.214	0.188	0.333	0.179
m63	0.001	0.092	-0.034	0.096	-0.033	0.096
malepat	-0.400	0.065	-0.424	0.063	-0.429	0.063
blackpat	-0.441	0.131	-0.369	0.131	-0.399	0.125

(a) and (c) based on all data: N=3889 patients; K=57 medical practices; (a) $\hat{\rho} = 0.154$, (c) $\hat{\rho} = 0.155$.
(b) cluster 5 deleted: N=3698 patients; K=56 medical practices; $\hat{\rho} = 0.171$.
robust standard errors are given in (a).

Table 4.2 *Summary of cluster deletion diagnostics from GEE fit and REGEE weights for selected clusters from the Medical Practice data.*

cluster	n_i	$tr(H_i)$	$DCLS_i$	weight
5	191	1.26	1.25	.32
19	197	0.87	0.36	.59
15	158	0.58	0.33	.80
29	89	0.32	0.27	.93

Table 4.3 *Parameter estimates and standard errors for the subset of Medical Practice data using (a) GEE, (b) Mallows observation-downweighting REGEE, and (c) Schweppe observation-downweighting REGEE.*

	(a)		(b)		(c)	
	$\hat{\beta}_j$	se	$\hat{\beta}_j$	se	$\hat{\beta}_j$	se
intercept	0.594	0.324	0.495	0.316	0.537	0.319
speclty	0.266	0.396	0.361	0.390	0.349	0.387
mdage	-0.326	0.207	-0.336	0.204	-0.323	0.206
m65	-0.044	0.408	0.031	0.401	0.005	0.415
patage	-0.068	0.165	-0.053	0.173	-0.062	0.161
noinsur	0.672	0.586	0.755	0.594	0.776	0.610
nbrmds	-0.263	0.124	-0.228	0.120	-0.265	0.131
m3	0.004	0.332	-0.010	0.351	-0.045	0.337
m63	-0.290	0.282	-0.258	0.284	-0.308	0.291
malepat	-1.342	0.392	-1.373	0.392	-1.369	0.401
blackpat	-0.602	0.351	-0.540	0.354	-0.609	0.346

All entries based on N = 171 patients in K=57 medical practices;
exchangeable correlation estimates are (a) $\hat{\rho} = .095$; (b) $\hat{\rho} = .043$; (c) $\hat{\rho} = .050$.

Table 4.4 *Summary of observation deletion diagnostics from GEE fit and REGEE weights for selected observations from the subset of Medical Practice data.*

						Weight	
cluster	obs. id	n_i	h_{it}	r_{it}	$DOBS_{it}$	Mallows	Schweppe
2	4	4	.142	-1.36	.0481	.65	.78
6	17	4	.097	-1.99	.0386	.80	.59
56	166	2	.046	2.69	.0328	.93	.41
55	164	3	.045	2.49	.0302	.95	.49

CHAPTER V

EFFICIENCY AND BIAS OF REGEE

5.1 Introduction

The previous chapter introduced REGEE, gave its asymptotic properties in Theorem 3, and demonstrated its resistance to influential observations through an example of medical practice data. This chapter considers the efficiency of REGEE through analytical and numerical means. In addition, robustness properties of REGEE are established numerically by simulating data from contaminated models. First, in section 5.2, justification for REGEE as a robust estimating equation procedure is provided while showing that it necessarily has suboptimal properties, at least for binary responses, and the Mallows class, in general. The conclusion is reached that efficiency must be sacrificed in order to gain robustness. The loss in efficiency is captured by the asymptotic relative efficiency (ARE) of REGEE with respect to GEE. The ARE is defined in this way because REGEE is proposed as an alternative methodology to GEE, so there is little interest in the efficiency of REGEE with respect to maximum likelihood. Additionally, as in GEE, the full likelihood is not always specified in the REGEE approach. In section 5.3, the individual roles of within cluster correlation of the responses, cluster size and design are described for special cases. In general, however, the roles are difficult to separate. Accordingly, for correlated binary responses, the ARE for certain designs is evaluated for a simple logit model with known parameters with the aid of a SAS IML computer program. Section 5.4 extends the investigation of the efficiency of Schweppe observation-downweighting REGEE for small sample sizes by

evaluating the ARE at estimated values of β obtained through repeated sampling from this model. Lastly, in section 5.5, similar simulation methods are employed to examine robustness by generating data from contaminated mixture models (Copas,1988). For the logit model and clusters of size two, the bias of GEE and Schweppe observation-downweighting REGEE is compared for different levels of contamination and different values of the tuning constant. In this way, the robustness of REGEE is demonstrated in its having smaller bias than GEE.

5.2 In Search of an Optimal Solution

The theory of estimating functions provides further perspective on the asymptotic properties of REGEE given by (4.1). Consider the non-robust elementary estimating functions, $g_{it} = Y_{it} - \mu_{it}(\beta)$, $t = 1, \ldots, n_i$, and $i = 1, \ldots, K$. The theory of Godambe and Heyde (1987) and Morton (1981) states that the optimal estimating function is given by

$$\mathrm{EF} = \sum_{i=1}^{K} C_i' V_i^{-1} g_i$$

where $g_i = (g_{i1}, \cdots, g_{in_i})'$, $C_i = \mathrm{E}_\beta[\partial g_i / \partial \beta]$ and $V_i = \mathrm{Var}(g_i)$. The solution, $\widehat{\beta}$ of the estimating equation, EF $= 0$, is optimal in the sense that it has the smallest asymptotic variance among all estimating functions that are linear combinations of the elementary estimating functions. Its variance is said to be smaller if $\mathrm{Var}(\widetilde{\beta}) - \mathrm{Var}(\widehat{\beta})$ is positive definite for all possible solutions $\widetilde{\beta}$ in the class of estimating functions considered. Thus GEE given by (3.1) which has the form EF $= 0$ is optimal. It follows that Mallows REGEE is suboptimal because it is in the same class of estimating functions as GEE, i.e, those that can be expressed as linear functions of $Y_i, i = 1, \ldots, K$. The question of Schweppe REGEE is not as straightforward because these functions are not linear in Y_i. Intuitively, one might expect a loss in efficiency for a gain in robustness. Indeed, this is the case as we now prove. Consider Schweppe observation-downweighting REGEE.

The robust elementary estimating function is $g_{it}^* = w_{it}(r_{it}, a)(Y_{it} - \mu_{it}) - c_{it}$, where w is a smooth weight function with tuning constant a, r_{it} is the Pearson residual, and $c_{it} = \mathrm{E}[w_{it}(r_{it}, a)(Y_{it} - \mu_{it})]$. By the above theory, the optimal estimating equation is

$$\mathrm{EF}^* = \sum_{i=1}^{K} C_i{}^{*\prime}(V_i{}^*)^{-1} g_i^*$$

where $g_i{}^* = (g_{i1}^*, \ldots, g_{in_i}^*)'$, $C_i^* = \mathrm{E}_\beta[\partial g_i^* / \partial \beta]$ and $V_i^* = \mathrm{Var}(g_i^*)$. When the optimality theory is applied to correlated binary responses the result is, surprisingly, $\mathrm{EF}^* = \mathrm{EF}$ (see Appendix 6). Thus in "optimizing" the robust elementary estimating function, g_i^*, the robustness is lost. Equation (4.1) is essentially of the form

$$\sum_{i=1}^{K} C_i' V_i^{-1} g_i^*,$$

which is necessarily less efficient than EF. Robustness can only be gained at the cost of efficiency, both of which are determined by w and a in REGEE. Next, the nature and magnitude of the effiiciency loss of REGEE compared to GEE is examined.

5.3 Asymptotic Relative Efficiency of REGEE to GEE

The REGEE procedure provides protection in the form of robustness against observations or clusters which deviate from the model. When all the data follows (3.1), i.e., when the model is not contaminated, the discussion in the previous section suggests that REGEE will be less efficient than GEE. This section defines and examines the asymptotic relative efficiency of REGEE to GEE. Consideration is limited to a working exchangeable correlation matrix, R_i, that is assumed to be correct, i.e., $R_i = corr(Y_i)$. If $J = 1\,1'$ is a $n_i \times n_i$ matrix of ones, then

$$corr(Y_i) = [1 + (n_i - 1)\rho]\tfrac{1}{n_i} J + (1 - \rho)(I - \tfrac{1}{n_i} J).$$

The corresponding inverse is

$$corr^{-1}(Y_i) = [1 + (n_i - 1)\rho]^{-1} \frac{1}{n_i} J + (1 - \rho)^{-1}(I - \frac{1}{n_i} J). \tag{5.1}$$

The efficiency loss of REGEE is given by the generalized asymptotic relative efficiency of $\widehat{\beta}_R$ to $\widehat{\beta}_G$ which is defined to be

$$ARE_{R:G} = |\text{var}(\widehat{\beta}_G)\text{var}^{-1}(\widehat{\beta}_R)|^{1/p} \tag{5.2}$$

Note that

$$ARE_{R:G} = \Big(\prod_{j=1}^{p}\lambda_j\Big)^{1/p}$$

where λ_j is the j-th eigenvalue of $\text{var}(\widehat{\beta}_G)\text{var}^{-1}(\widehat{\beta}_R)$. Because (5.2) is generally complicated, special cases when cluster sizes are equal are considered. The sandwich form of the asymptotic variance matrix of $\widehat{\beta}_G$ is given by

$$\text{var}(\widehat{\beta}_G) = (\sum_{i=1}^{K} D_i' V_i^{-1} D_i)^{-1}\phi. \tag{5.3}$$

Let W_i denote the diagonal matrix of GEE weights, $w_{it} = (\partial\mu_{it}/\partial\eta_{it})\text{v}_{it}^{1/2}(\mu_{it})$, which is not the W_i in either of the previous two chapters. An equivalent expression for (5.3) is

$$\text{var}(\widehat{\beta}_G) = (\sum_{i=1}^{K} \text{X}_i' W_i R_i^{-1} W_i \text{X}_i)^{-1}\phi. \tag{5.4}$$

The weights will be constant, $w_{it} = w$, for the identity link function with constant variance (i.e., for the linear model), and more generally, for generalized linear models with link equal to the variance stabilizing transformation (McCullagh & Nelder, 1989). Substituting (5.1) into (5.4) yields

$$\text{var}(\widehat{\beta}_G) = w^{-2}(e_n^{-1}SS_{uc} + f^{-1}SS_{wc})^{-1}\phi \tag{5.5}$$

where $e_n = 1 + (n-1)\rho$, $f = 1 - \rho$, and SS_{uc} and SS_{wc} are the between-cluster (uncorrected) and within cluster sum of squares and cross products matrices of x. Specifically,

$$SS_{uc} = \sum_{i=1}^{K}\tfrac{1}{n}X_i' J X_i \qquad SS_{wc} = \sum_{i=1}^{K} X_i' X_i - \sum_{i=1}^{K}\tfrac{1}{n}X_i' J X_i$$

In particular, for p = 1,

$$SS_{uc} = n\sum_{i=1}^{K}\overline{x}_{i.}^2, \text{ and } \quad SS_{wc} = \sum_{i=1}^{K}\sum_{t=1}^{n}x_{it}^2 - n\sum_{i=1}^{K}\overline{x}_{i.}^2$$

where $\overline{x}_{i.}$ is the mean of the x_{it} in the t-th cluster. If all the covariates are cluster level, $SS_{wc} = 0$. If covariates vary within cluster and are mean balanced at zero, i.e.,$\overline{x}_{i.} = 0$, $i = 1, ..., K$, then $SS_{uc} = 0$.

Next, assuming as in Chapter 4 that var(ψ_i) $= \Gamma_i$var(Y_i)Γ_i, the asymptotic variance matrix of the REGEE estimate is

$$\text{var}(\widehat{\beta}_R) = H_1^{-1}H_2H_1^{-1}\phi, \text{ where} \tag{5.6}$$

$$H_1 = (\sum_{i=1}^{k}D_i'V_i^{-1}\Gamma_iD_i), \text{ and } H_2 = (\sum_{i=1}^{k}D_i'V_i^{-1}\Gamma_iV_i\Gamma_iV_i^{-1}D_i).$$

In the next section, the role of the exchangeable correlation, ρ, the common cluster size, n, and the design, X in (5.2) is considered for special cases.

5.3.1 ARE when all the covariates are cluster-level

If all covariates are cluster level, $\Gamma_i = b_iI$, where b_i is a scalar quantity. This follows for Mallows cluster-downweighting because all observations receive equal weight by definition. For observation downweighting, Mallows or Schweppe, $b_{it} = b_i$ follows from $\mu_{it} = \mu_i$. Similarly, cluster level covariates imply $w_{it} = w_i$, and in (5.6),

$$H_1 = (\sum_{i=1}^{k}b_iw_i^2\mathrm{X}_i'\mathrm{R}_i^{-1}\mathrm{X}_i) = e_n^{-1}SS_{uc,}^{(12)}$$

$$H_2 = (\sum_{i=1}^{k}b_i^2w_i^2\mathrm{X}_i'\mathrm{R}_i^{-1}\mathrm{X}_i) = e_n^{-1}SS_{uc}^{(22)}.$$

where,

$$SS_{uc}^{(12)} = \sum_{i=1}^{K}\tfrac{1}{n}b_iw_i^2X_i'JX_i, \text{ and } SS_{uc}^{(22)} = \sum_{i=1}^{K}\tfrac{1}{n}b_i^2w_i^2X_i'JX_i.$$

Expression (5.2) simplifies to

$$ARE_{R:G} = |SS_{uc}^{(02)\,-1}SS_{uc}^{(12)}SS^{(22)-1}SS_{uc}^{(12)}|^{1/p}, \tag{5.7}$$

where,

$$SS_{uc}^{(02)} = \sum_{i=1}^{K}\tfrac{1}{n}w_i^2X_i'JX_i.$$

In summary, if all the covariates are cluster level, the ARE is a function of the REGEE weights, but not of the common cluster size. It is a function of the correlation only through the weights. This holds for any generalized linear model, such as the logit link and binomial variance function for correlated binary outcomes.

5.3.2 ARE when covariates vary within cluster

In general, when covariates vary within cluster, (5.2) is complicated and depends upon the correlation and the common cluster size. A special case is for constant GEE weights, i.e., $w_{it} = w$ in the context of Mallows cluster-downweighting which implies, $\Gamma_i = b_i I$, as noted above. Then,

$$H_1 = (w^2 \sum_{i=1}^{k} b_i \mathbf{X}_i' \mathbf{R}_i^{-1} \mathbf{X}_i) = w^2 (e_n^{-1} SS_{uc}^{(10)} + f^{-1} SS_{wc}^{(10)})$$
$$H_2 = (w^2 \sum_{i=1}^{k} b_i^2 \mathbf{X}_i' \mathbf{R}_i^{-1} \mathbf{X}_i) = w^2 (e_n^{-1} SS_{uc}^{(20)} + f^{-1} SS_{wc}^{(10)}).$$

where,

$$SS_{uc}^{(10)} = \sum_{i=1}^{K} \tfrac{1}{n} b_i X_i' J X_i, \qquad SS_{uc}^{(20)} = \sum_{i=1}^{K} \tfrac{1}{n} b_i^2 X_i' J X_i.$$
$$SS_{wc}^{(10)} = \sum_{i=1}^{K} b_i X_i' X_i - \sum_{i=1}^{K} \tfrac{1}{n} b_i X_i' J X_i, \text{ and } SS_{wc}^{(20)} = \sum_{i=1}^{K} b_i^2 X_i' X_i - \sum_{i=1}^{K} \tfrac{1}{n} b_i^2 X_i' J X_i$$

If, additionally, the covariates are mean balanced at zero, $SS_{uc} = SS_{uc}^{(10)} = SS_{uc}^{(20)} = 0$, then from (5.2) and (5.5),

$$ARE_{R:G} = |SS_{wc}^{-1} SS_{wc}^{(10)} SS_{wc}^{(20)-1} SS_{wc}^{(10)}|^{1/p}.$$

In other words, in the Mallows cluster-downweighting class, for constant GEE weights, equal cluster sizes, and covariates that vary within clusters but are mean-balanced, the generalized asymptotic relative efficiency of $\widehat{\beta}_R$ to $\widehat{\beta}_G$ depends on the tuning constant through the b_i, but does not depend on n. It depends on ρ only through the REGEE weights. For observation downweighting REGEE, Mallows or Schweppe, the b_{it} vary

within clusters when covariates are not all cluster level, and so the asymptotic relative efficiency will depend on cluster size and correlation.

For nonconstant weights, the matrices resulting from reexpressing (5.4) and (5.6) are difficult to interpret because they depend not only on the design, but also on the values of ρ and β. When cluster sizes vary, the asymptotic variance of $\widehat{\beta}_G$ will, generally, depend on the correlation, average cluster size, and coefficient of variation of the cluster sizes (Mancl & Leroux, 1995).

The ARE of the parameter estimates excluding the intercept is often of interest, instead of (5.2). An expression like (5.5) is given by Mancl & Leroux (1995) with SS_c replaced by

$$\sum_{i=1}^{K} \frac{1}{n} X_i' J X_i - \frac{1}{nK} (\sum_{j=1}^{K} X_j' 1)(\sum_{l=1}^{K} 1' X_l).$$

This result is obtained by applying a standard matrix algebra result for the inverse of a partitioned matrix. For REGEE, however, the analagous formula for (5.6) is algebraically complex. As illustrated in the next section, however, the conclusions drawn for the entire regression parameter appear to apply to the parameter vector which omits the intercept.

5.3.3 Computational approach for correlated binary responses

In order to further study the roles of n, ρ, and the design, in the asymptotic relative efficiency of REGEE to GEE, (5.2) was determined for eight different scenarios by applying an IML computer program. Throughout the rest of the Chapter, consideration is given to a model for correlated binary responses which has a logit link and binomial variance function. It is assumed that the model has a single covariate, $p = 1$, which was either constant within cluster or varying within cluster but mean-balanced with $\overline{x}_{i.} = 0$. Furthermore, we assume the intercept parameter is α = -2.0 and the covariate parameter is $\beta = 0.8$. Primary interest is in the asymptotic relative

efficiency of $\hat{\beta}$ defined by $ARE_\beta =$ var($\hat{\beta}_G$)/var($\hat{\beta}_R$). For comparative purposes, $K = 50$ throughout, and n, ρ, and the design, are varied.

Four designs are considered and they are summarized in Table 5.1. They will be referred to as designs A, B, C, and D respectively. In each design, the common cluster size is either 2 or 4. In designs A and C, the covariate was constant within cluster, and in designs B and D, it varied within cluster.

Table 5.1 *Designs, X_i, considered for ARE_β*

	cluster level $\delta_i = -1 + \frac{2(i-1)}{k-1}$	vary within cluster $\Delta_i = i/K$
$n = 2$	*Design A* $\begin{pmatrix} 1 & \delta_i \\ 1 & \delta_i \end{pmatrix}$	*Design B* $\begin{pmatrix} 1 & \Delta_i \\ 1 & -\Delta_i \end{pmatrix}$
$n = 4$	*Design C* $\begin{pmatrix} 1 & \delta_i \\ 1 & \delta_i \\ 1 & \delta_i \\ 1 & \delta_i \end{pmatrix}$	*Design D* $\begin{pmatrix} 1 & \Delta_i \\ 1 & \Delta_i/3 \\ 1 & -\Delta_i/3 \\ 1 & -\Delta_i \end{pmatrix}$

For each design, ARE_β was evaluated for $\rho = .3$ and $\rho = .7$ giving a total of eight scenarios for Schweppe observation-downweighting (Table 5.2), Mallows observation-downweighting (Table 5.3) and Mallows cluster-downweighting (Table 5.4). For each scenario in each table, the asymptotic relative efficiency is given for a range of tuning constants, a, applied to the weight function, $w_{it}(v_{it}, a) = exp\{-(v_{it}/a)^2\}$, where v_{it} is specified later for each class. Also provided for each scenario is a measure, $ratio_b = max(b_{it})/min(b_{it})$, evaluated over $t = 1, ..., n$, and $i = 1, ... K$. It is a crude measure of loss of efficiency, since larger values indicate a wider range of nonoptimal weights applied to the observations. Finally, the eigenvalues of var($\hat{\alpha}_G, \hat{\beta}_G$)var^{-1}($\hat{\alpha}_R, \hat{\beta}_R$), denoted λ_1 for the largest and λ_2 for the smallest are given

for each scenario. The eigenvalue, λ_1, corresponds to the maximum $\text{ARE}_{R:G}$ obtained among all linear combinations of α and β, and λ_2 corresponds to the minimum $\text{ARE}_{R:G}$, and together they determine $\text{ARE}_{R:G}$ as indicated earlier. As expected by arguments given in section 5.2 , $\lambda_1 \leq 1$. Additionally, it was found that ARE_β, necessarily between λ_1 and λ_2, was generally closer to λ_2. Note throughout that as a approached infinity which represents GEE, the weights and thus $ratio_b$, ARE_β, and $\text{ARE}_{R:G}$ approached 1.

Results for Schweppe observation downweighting

The downweighting function was applied to the Pearson residuals, i.e., $v_{it} = r_{it}$. For the designs in which the covariate was cluster-level, Table 5.2 shows that the results were identical, indicating that ARE_β does not depend on correlation or common cluster size. REGEE was less efficient when the covariate varied within cluster. For these designs the ARE_β depended on sample size and correlation, and in particular REGEE was less efficient when $\rho = .7$ than when $\rho = .3$. Interestingly, for all eight scenarios considered, the greatest loss in efficiency occurred at approximately $a = 1.75$. Similarly, $ratio_b$, attained its greatest value at 1.75. Figure 5.1 plots ARE_β and $ratio_b$ versus a, for the cluster-level designs, A and C. As a became smaller the efficiency actually increased due to excessive downweighting resulting in a loss of discriminating power to detect the data which was the most influential. Since $b_{it} = b_i(\mu_{it})$ is a monotonic function of the mean only, $ratio_b$ is the same for every design considered since $\min(\mu_{it}) = 0.06$ and $\max(\mu_{it}) = 0.23$. In general, $ratio_b$ indicates that the loss of efficiency increases with increasing variation in the covariates, X, and with increasing magnitude of β in absolute terms. The relationship between $ratio_b$ and efficiency for the cluster level designs is also illustrated by Figure 5.2. This shows that two very different tuning constants may give very similar efficiency, but one represents excessive downweighting, and the other does not. Both figures show that as robustness increases, efficiency decreases. The

relationship of efficiency and tuning constant is illustrated for design B in Figure 5.3 and for design D in Figure 5.4.

Results for Mallows observation downweighting

For the eight scenarios, we applied the weight function to the observation leverages by setting $v_{it} = h_{it}$, and considered a range of tuning constants given by $a = mp/N, m = .5$, 1, 1.5, 2, 2.5, 3, 4, and 5, where p/N is the average of h_{it} over all observations. Generally, efficiency increased as a increased. For the designs in which the covariate was cluster-level, Table 5.3 shows that the results were identical, indicating that ARE_β does not depend on correlation or common cluster size. For designs in which the covariate varied within cluster, ARE_β depended on n and ρ. For design D ($n = 4$), REGEE was less efficient when $\rho = .7$ than when $\rho = .3$. However, for design B ($n = 2$), ARE_β was very similar for the different correlations. The tuning constant of 3 times the average of the observation leverages, which was applied to the subset of medical practice data in section 4.3.2, resulted in high efficiency for designs A, B, and C, but somewhat lower for D.

Results for Mallows cluster downweighting

For the eight scenarios, we applied the weight function to the cluster leverages by setting $v_{it} = H_i$, and considered a range of tuning constants given by $.04m$, for the range of m given above. Note that for all designs, the average of the H_i is $np/N = .04$. For the designs in which the covariate was cluster-level, Table 5.4 shows that the results were the same and, furthermore, they were identical to the corresponding result in Table 5.3. On the other hand, for designs in which the covariate varied within cluster, the ARE depended on n and ρ, but varied little.

5.4 Efficiency of REGEE to GEE for Small Sample Sizes

The efficiency for small sample sizes may be quite different than the asymptotic relative efficiency considered in the previous section. Repeated samples (y_{i1}, y_{i2}) from the simple logit model (α = -2, $\beta = 0.8$) described in section 5.3.3 were generated in order to assess the efficiency of Schweppe observation downweighting REGEE with respect to GEE for designs A and B in Table 5.1. For $K = 50$, 100 and 200, and for $\rho = .3$ and $\rho = .7$, the behavior of the variance estimators of $\widehat{\beta}_G$ and $\widehat{\beta}_R$ is studied rather than the variances themselves as in (5.2). Parameter estimates were obtained for each of 1000 simulations in which the REGEE algorithm converged in 100 iterations or less. The estimated relative efficiency ($Eff_{\beta,K}$) given in Table 5.5 is evaluated as the ratio of the estimated mean squared error (MSE) of the $\widehat{\beta}_G$'s from GEE to the average of the MSE of the $\widehat{\beta}_R$'s from REGEE, after the upper and lower 5% of each set were omitted. The MSE is calculated as the sum of the estimated bias and the sample variance of the $\widehat{\beta}$'s,

$$\text{MSE}(\hat{\beta}) = \left\{\beta - \{average(\widehat{\beta})\}\right\}^2 + var(\widehat{\beta}).$$

The MSE was used instead of the sample variance because there was some nonneglibale bias for GEE and REGEE as reported in the next section. For comparative purposes, ARE_β from Table 5.2 is also given.

The small sample efficiency, $Eff_{\beta,\,50}$, from the simulation study reported in Table 5.5 for $a = 3$, and Table 5.6 for $a = 2.25$, is consistently less than the corresponding asymptotic relative efficiency, ARE_β. Thus, for $K = 50$ and for the designs considered, ARE_β is not a reliable indicator of the actual efficiency. As expected, as K increases, the finite sample efficiency, Eff_β increases. Tables 5.5 and 5.6 show that, in general, for correlated binary responses, the efficiency of Schweppe observation downweighting REGEE to GEE is lower for larger values of correlation, although the opposite was observed for $K = 200$ for Design B.

5.5 Small Sample Bias of REGEE and GEE under Models of Contamination

In order to illustrate the robustness of REGEE, data was generated from models that have a certain level of contamination, ϵ. In particular, contamination models, (2.35), are specified such that each binary observation has mean μ_{it} with probability $(1-\epsilon)$ and mean $1-\mu_{it}$ with probability ϵ, where μ_{it} is determined from the model and designs of the previous section. The data (y_{i1}, y_{i2}) are generated at random from the corresponding multinomial distribution specified by the correlation, specified a priori, and the two randomly selected means. The performance of GEE and Schweppe observation-downweighting REGEE with $a = 3$ and $a = 2.25$, are evaluated for $\epsilon = 0$, 0.03, 0.05, and 0.10. The robustness of each procedure is interpreted in terms of the estimated bias, as described in the previous section, based on 1000 simulations per scenario. Table 5.7 reports the bias for $K = 50$, except where noted. In particular, there was non-negligible bias for $\widehat{\beta}_G$ under a model without contamination (0%), and the bias was larger for $\widehat{\beta}_R$. The bias decreased when K increased from 50 to 100, but was still larger for REGEE. The bias at $K = 100$ is shown to illustrate that as K gets larger the bias should approach 0, the asymptotic bias for both procedures. Under models with contamination, the bias of REGEE estimates is lower than bias of GEE estimates, illustrating the robustness of REGEE. For all scenarios, the bias of REGEE is about half the bias of GEE when the contamination is at 3%. For 5% contamination there is some improvement of REGEE over GEE, but if the contamination is as large as 10%, then the REGEE procedure is not any more resistant than GEE. REGEE with $a = 2.25$ gives smaller bias than REGEE with $a = 3.0$, but the gain in robustness illustrated by the smaller bias is not a great improvement in consideration of the loss in efficiency indicated in Table 5.2. For example, for design A, the bias improves by only .004 (or .005), whereas ARE_β decreases from .894 to .791. For the designs considered here, then, a

tuning constant of $a = 3$ is recommended over $a = 2.25$. For designs with a cluster level covariate the bias is greater for $\rho = .7$ when $\epsilon = 0$ but greater for $\rho = .3$ when $\epsilon = 0.03$. For designs with a covariate that varies within cluster the relationships are reversed.

Table 5.2 *The Asymptotic Relative Efficiency of Schweppe observation downwieghting REGEE to GEE for correlated binary responses*

a	5	4	3.5	3	2.5	2.25	2	1.75	1.5	1	0.5
ratio	1.65	2.16	2.68	3.62	5.39	6.62	7.65	7.75	6.70	3.69	1.67
Cluster level designs, A and C											
ARE_β	.982	.959	.935	.894	.829	.791	.758	.747	.774	.890	.982
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	.981	.956	.929	.883	.809	.766	.730	.722	.759	.887	.977
Design B, $\rho = .3$											
ARE_β	.968	.929	.889	.825	.733	.684	.645	.637	.675	.825	.968
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	.968	.928	.887	.823	.729	.678	.640	.632	.671	.822	.967
Design B, $\rho = .7$											
ARE_β	.916	.824	.742	.629	.498	.439	.400	.395	.439	.634	.920
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	.906	.805	.717	.599	.465	.407	.368	.362	.403	.598	.913
Design D, $\rho = .3$											
ARE_β	.967	.930	.893	.837	.758	.715	.680	.664	.685	.821	.970
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	.967	.929	.892	.835	.756	.713	.677	.662	.682	.817	.969
Design D, $\rho = .7$											
ARE_β	.922	.840	.769	.673	.558	.504	.463	.446	.467	.642	.933
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	.910	.819	.741	.639	.521	.467	.425	.408	.428	.603	.924

a = tuning constant which is applied to weight function, $exp\{-(r_{it}/a)^2\}$
$ratio = max(b_{it})/min(b_{it})$
$ARE_\beta = \text{var}(\widehat{\beta}_G)/\text{var}(\widehat{\beta}_R)$
λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\widehat{\alpha}_G,\widehat{\beta}_G)\text{var}^{-1}(\widehat{\alpha}_R,\widehat{\beta}_R)$

Figure 5.1 ARE_β and $ratio_b$ versus tuning constant for cluster level designs

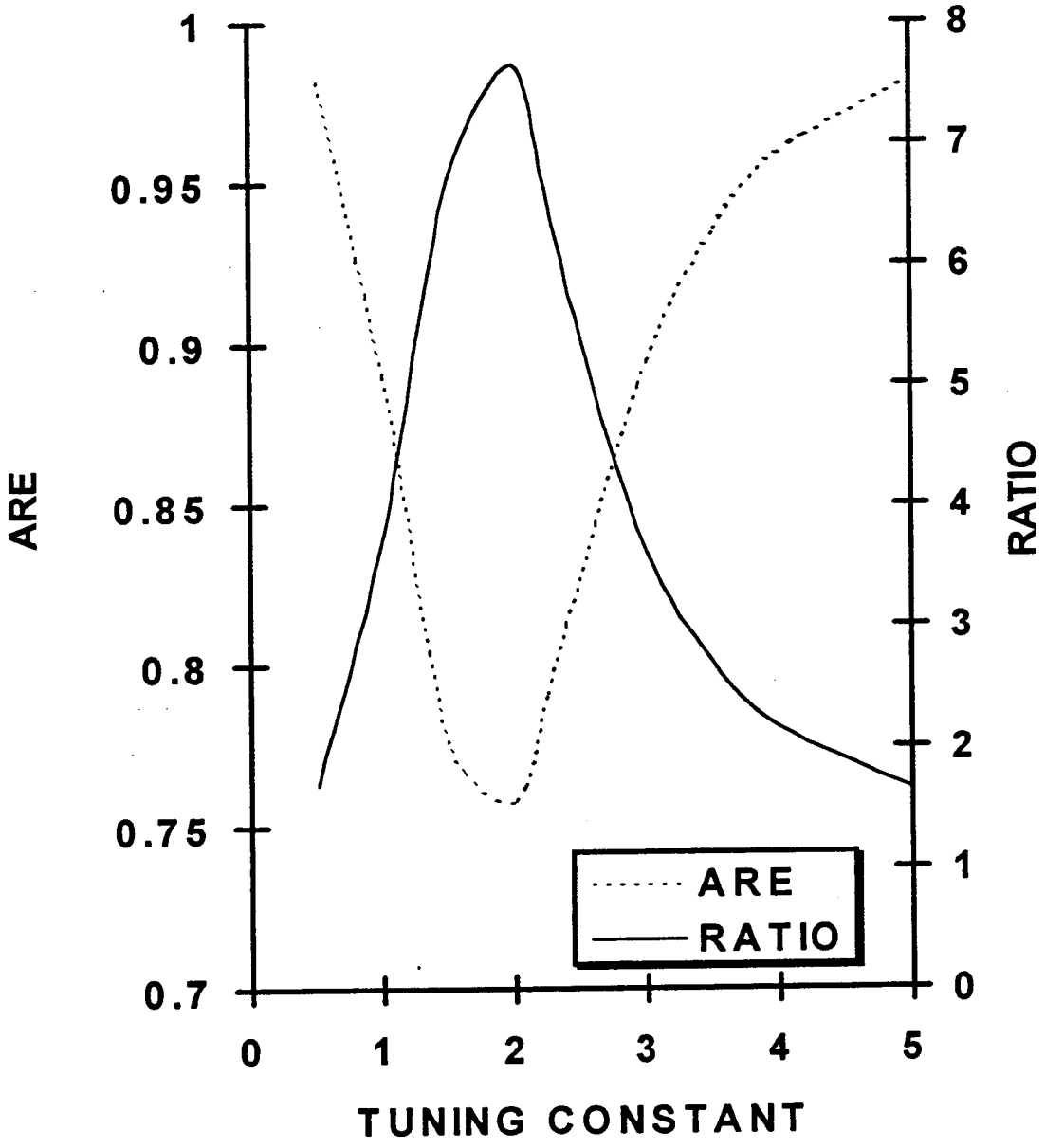


Figure 5.2 ARE_β versus $ratio_b$ for cluster level designs

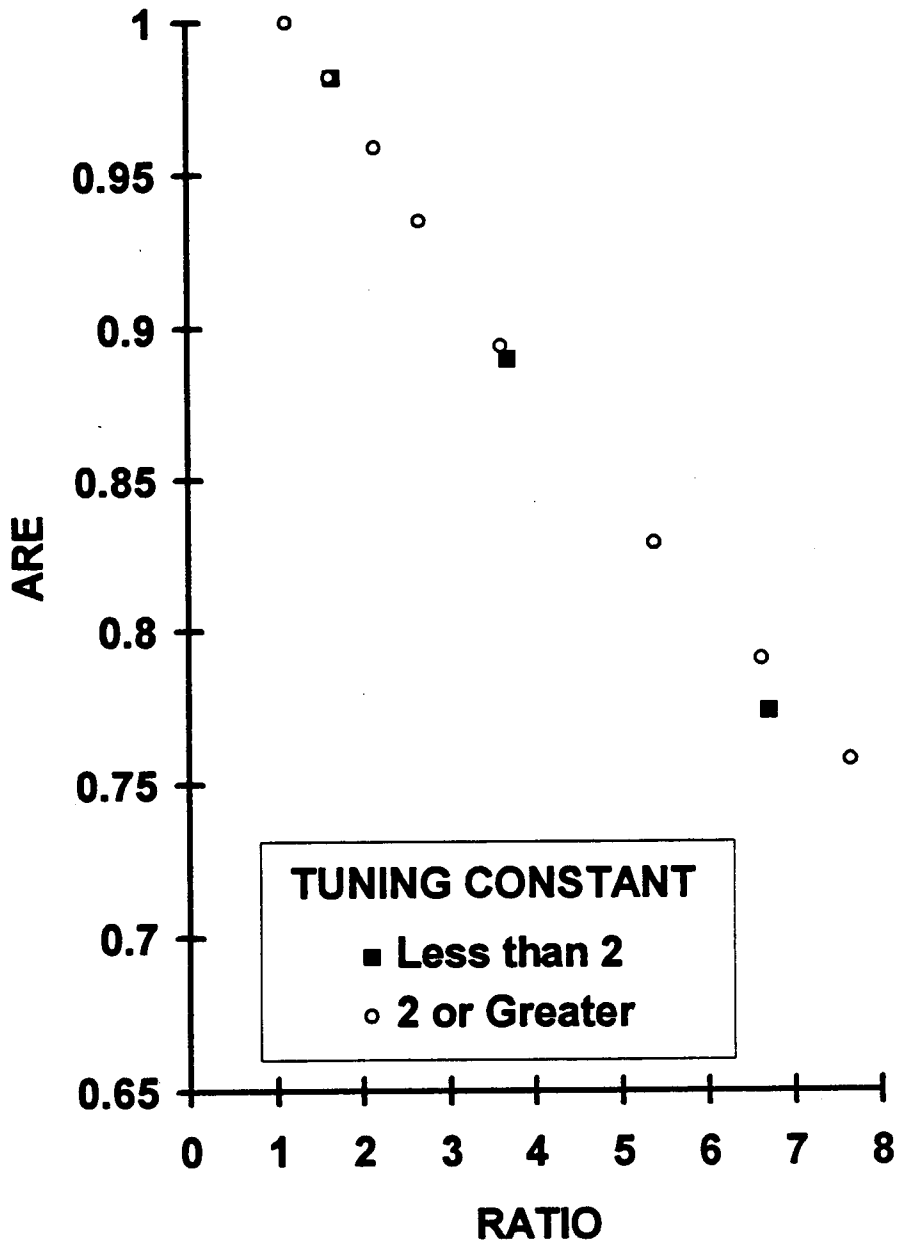


Figure 5.3 ARE_β versus tuning constant for Design B for $\rho = .3$ and $.7$

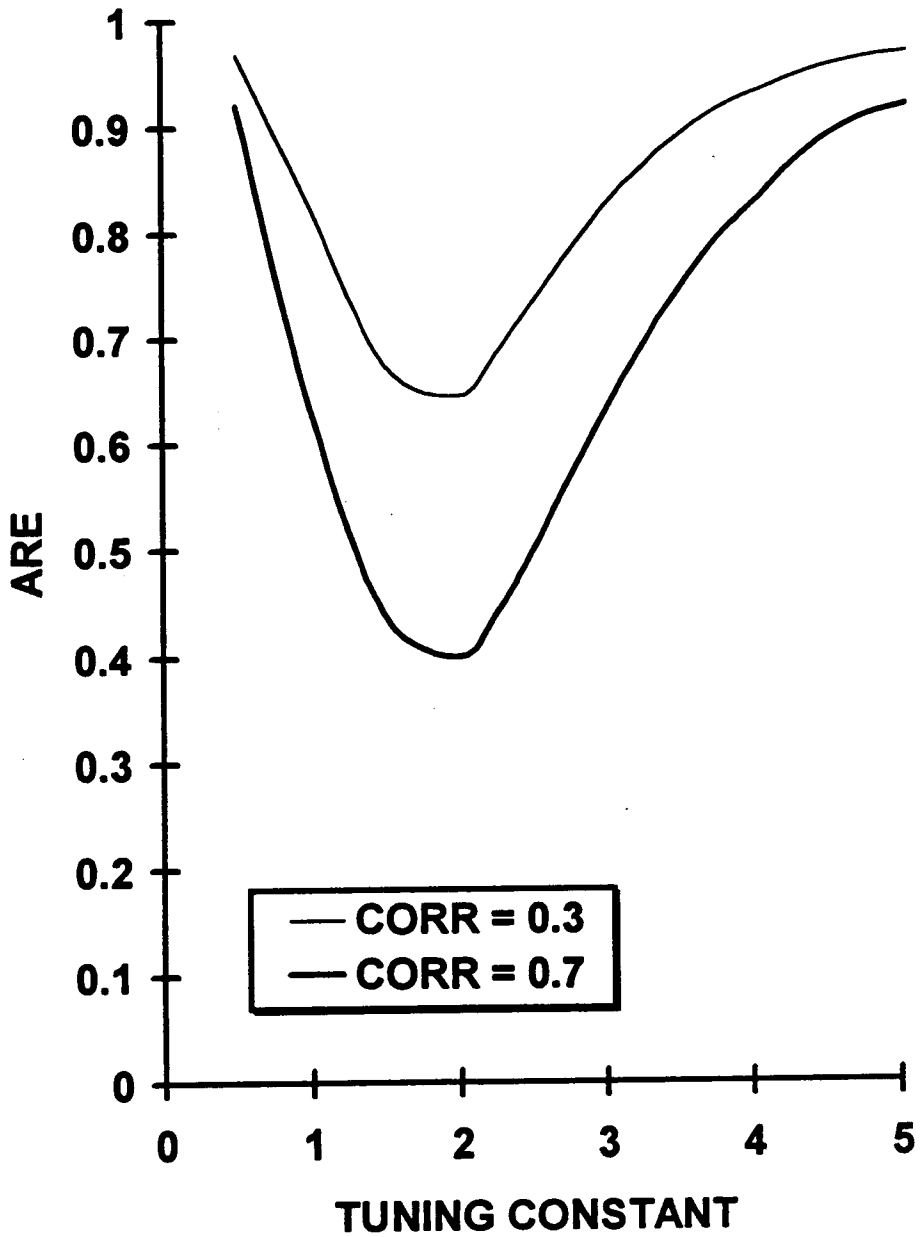


Figure 5.4 ARE_β versus tuning constant for Design D for $\rho = .3$ and $\rho = .7$

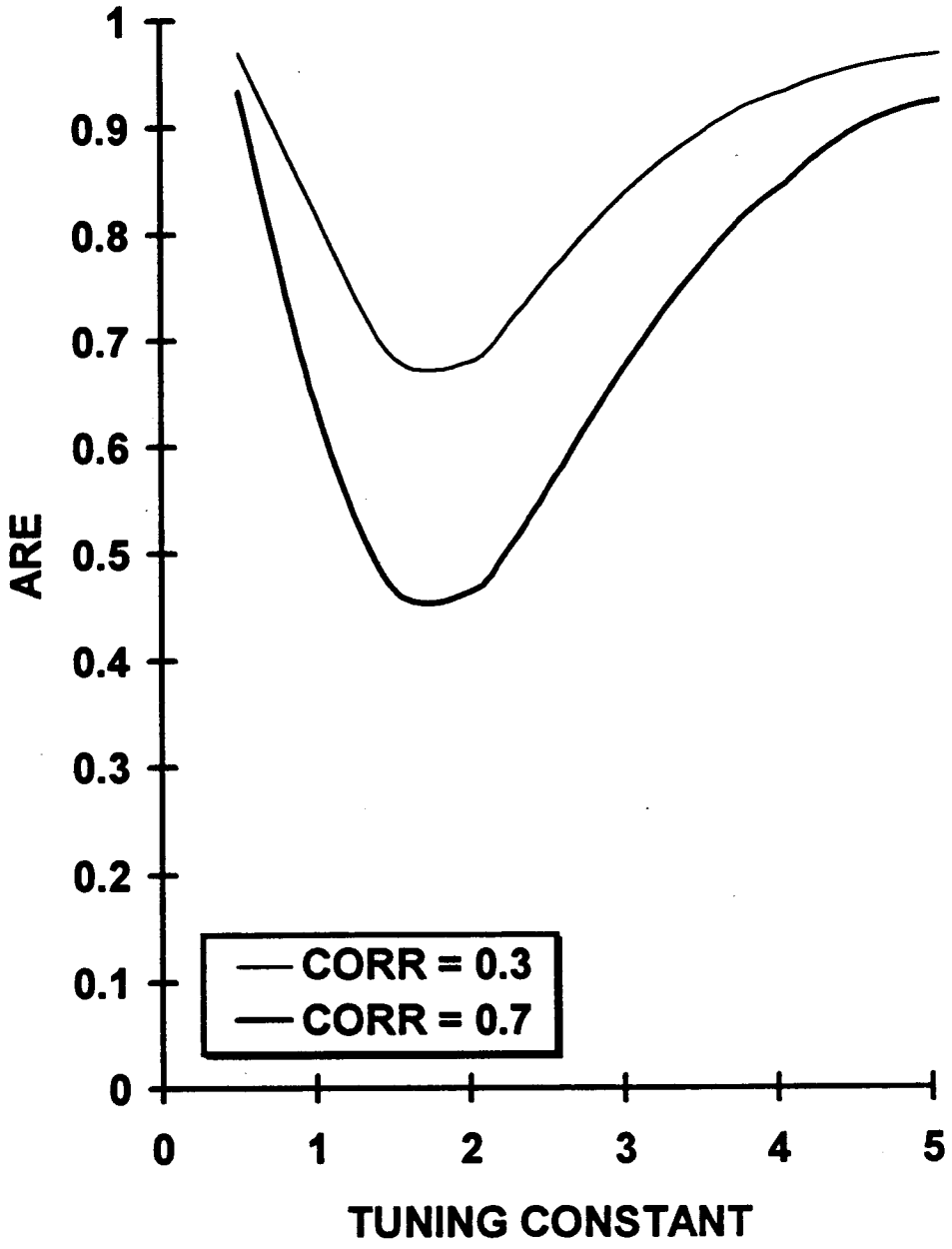


Table 5.3 *The Asymptotic Relative Efficiency of Mallows observation downweighting REGEE to GEE for correlated binary responses*

a_o	5	4	3	2.5	2	1.5	1	.5
Cluster level designs, A and C								
ratio	1.26	1.44	1.91	2.53	4.26	13.2	331	++
ARE_β	.997	.993	.978	.958	.912	.802	.565	.154
λ_1	1	1	1	1	1	1	.904	.515
λ_2	.994	.987	.961	.927	.852	.698	.448	.131
Design B, $\rho = .3$								
ratio	1.23	1.39	1.80	2.32	3.73	10.4	194	++
ARE_β	.997	.994	.982	.964	.923	.820	.574	.151
λ_1	1	.999	.997	.995	.987	.962	.846	.492
λ_2	.995	.989	.968	.939	.874	.731	.476	.133
Design B, $\rho = .7$								
ratio	1.21	1.34	1.69	2.13	3.26	8.18	113	++
ARE_β	.998	.994	.982	.966	.927	.828	.589	.155
λ_1	1	.999	.997	.993	.984	.954	.824	.480
λ_2	.995	.988	.964	.932	.862	.712	.463	.130
Design D, $\rho = .3$								
ratio	1.97	2.87	6.53	14.9	68.2	++	++	++
ARE_β	.966	.929	.838	.757	.647	.507	.322	.138
λ_1	1	1	1	1	1	.983	.919	.725
λ_2	.956	.908	.797	.703	.583	.446	.283	.124
Design D, $\rho = .7$								
ratio	1.84	2.59	5.42	11.4	44.8	863	++	++
ARE_β	.948	.892	.770	.673	.561	.450	.325	.136
λ_1	1	1	1	1	1	.991	.942	.694
λ_2	.930	.859	.710	.600	.480	.371	.261	.113

++ = greater than 1000
tuning constant is $a = a_o \times p/N$, where $p = 2$, and $N = 100$ or 200.
$ratio = max(b_{it})/min(b_{it})$
$ARE_\beta = \text{var}(\widehat{\beta}_G)/\text{var}(\widehat{\beta}_R)$
λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\widehat{\alpha}_G, \widehat{\beta}_G)\text{var}^{-1}(\widehat{\alpha}_R, \widehat{\beta}_R)$

Table 5.4 *The Asymptotic Relative Efficiency of Mallows cluster downweighting REGEE to GEE for correlated binary responses*

a	.20	.16	.12	.10	.08	.06	.04	.02
Cluster level designs, A and C								
ratio	1.26	1.44	1.91	2.53	4.26	13.2	331	++
ARE_β	.997	.993	.978	.958	.912	.802	.565	.154
λ_1	1	1	1	1	1	1	.904	.515
λ_2	.994	.987	.961	.927	.852	.698	.448	.131
Design B, $\rho = .3$								
ratio	1.15	1.24	1.46	1.73	2.36	4.59	30.8	++
ARE_β	.998	.996	.988	.976	.945	.850	.571	.155
λ_1	.999	.997	.991	.982	.961	.906	.768	.494
λ_2	.998	.996	.988	.975	.942	.842	.546	.135
Design B, $\rho = .7$								
ratio	1.15	1.24	1.46	1.72	2.34	4.54	30.1	++
ARE_β	.998	.996	.989	.977	.946	.855	.584	.163
λ_1	.999	.997	.990	.981	.960	.903	.759	.478
λ_2	.998	.996	.988	.976	.943	.844	.549	.136
Design D, $\rho = .3$								
ratio	1.15	1.25	1.49	1.77	2.44	4.89	35.5	++
ARE_β	.998	.996	.987	.974	.940	.838	.549	.145
λ_1	.998	.996	.989	.980	.956	.896	.749	.473
λ_2	.998	.996	.987	.973	.938	.832	.532	.132
Design D, $\rho = .7$								
ratio	1.15	1.25	1.49	1.77	2.44	4.90	35.7	++
ARE_β	.998	.996	.987	.974	.941	.842	.562	.155
λ_1	.998	.996	.989	.979	.955	.894	.743	.464
λ_2	.998	.996	.987	.973	.938	.832	.532	.133

++ = greater than 1000
a = tuning constant which is applied to weight function, $exp\{-(r_{it}/a)^2\}$
$ratio = max(b_{it})/min(b_{it})$
$ARE_\beta = \text{var}(\widehat{\beta}_G)/\text{var}(\widehat{\beta}_R)$
λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\widehat{\alpha}_G,\widehat{\beta}_G)\text{var}^{-1}(\widehat{\alpha}_R,\widehat{\beta}_R)$

Table 5.5 *Efficiency of Schweppe observation downweighting REGEE($a = 3$) to GEE for designs with two correlated binary responses*

Design	ρ	ARE_β	$Eff_{\beta,200}$	$Eff_{\beta,100}$	$Eff_{\beta,50}$
A	.3	.894	.850	.666	.590
A	.7	.894	.786	.665	.478
B	.3	.825	.687	.394	.497
B	.7	.629	.827	.202	.226

$Eff_{\beta,K} = \text{MSE}(\hat{\beta}_G)/\text{MSE}(\hat{\beta}_R)$ for designs with K clusters.

Table 5.6 *Efficiency of Schweppe observation downweighting REGEE ($a = 2.25$) to GEE for designs with two correlated binary responses*

Design	ρ	ARE_β	$Eff_{\beta,200}$	$Eff_{\beta,100}$	$Eff_{\beta,50}$
A	.3	.791	.768	.653	.612
A	.7	.791	.708	.614	.478
B	.3	.684	.562	.441	.506
B	.7	.439	.668	.189	.228

$Eff_{\beta,K} = \text{MSE}(\hat{\beta}_G)/\text{MSE}(\hat{\beta}_R)$ for designs with K clusters.

Table 5.7 *Estimated Bias of GEE and Schweppe observation downweighting REGEE for correlated binary responses based on 1000 simulations from a mixture model,* $\mathrm{E}[Y] = (1-\epsilon)\mu + \epsilon(1-\mu)$

ϵ, *percent contamination*	GEE	$a = 3$	$a = 2.25$	GEE	$a = 3$	$a = 2.25$
	Design A, $\rho = .3$			Design A, $\rho = .7$		
0 ($K = 200$)	.010	.016	.020	.009	.020	.023
0 ($K = 100$)	.022	.056	.052	.026	.066	.067
0	.055	.103	.082	.080	.184	.162
3	-.116	-.074	-.070	-.100	-.028	-.023
5	-.246	-.228	-.208	-.231	-.185	-.163
10	-.358	-.347	-.338	-.356	-.350	-.335
	Design B, $\rho = .3$			Design B, $\rho = .7$		
0 $K = 200$)	.005	.020	.028	-.113	-.066	-.042
0 ($K = 100$)	.069	.166	.129	-.070	.103	.115
0	.082	.144	.128	-.069	.101	.103
3	-.084	-.046	-.042	-.196	-.120	-.098
5	-.189	-.172	-.161	-.311	-.254	-.226
10	-.335	-.328	-.322	-.400	-.388	-.379

a = tuning constant which is applied to weight function, $exp\{-(r_{it}/a)^2\}$
K = 50 unless otherwise indicated

CHAPTER VI

SUMMARY AND FUTURE RESEARCH

6.1 Summary

Although the Generalized Estimating Equations procedure is applied widely in medical and biological sciences, the influence of observations and clusters has been overlooked in the statistical literature. This work addressed the problem of the influence that observations and clusters can have on regression parameter estimates and fitted values. In a classical approach, deletion diagnostics were proposed in Chapter 3 which estimate the effect of the deletion of an observation or a cluster. These deletion diagnostics are generalizations of DBETA of Belsley et al. and Cook's Distance of Cook (1977) in linear regression. Unlike the diagnostics for linear regression, however, the proposed diagnostics are approximations of their exact counterparts. Nevertheless, the diagnostics were shown to be good approximations. Furthermore, their computation by a SAS IML computer program was fast. In addition to these diagnostics for influence, diagnostics for the leverage of an observation and of a cluster were also proposed. Further work and experience is needed to increase understanding of the relationships these diagnostics, both influence and leverage, have with cluster size and correlation.

The second part of the work in Chapter 4 addressed the problem of influence through the introduction of a modification of the GEE procedure called Resistant Generalized Estimating Equations (REGEE). Unlike GEE, REGEE gives regression parameter estimates that are resistant to the influence of a small subset of the data. This is achieved by the automatic downweighting of influential observations in an iteratively reweighted least squares algorithm that is a modification of the GEE algorithm.

Although Theorem 3 in Chapter 4 indicates that REGEE applies to the same class of models considered by GEE, limitations as outlined in section 4.2.1 exist in its actual implementation. In the Mallows class, the same assumptions required in GEE are required, namely, the specification of the first two moments of the marginal distributions of the responses. For Schweppe observation downweighting, the full marginal univariate distributions are required. Except for bivariate binary or normal data, cluster downweighting REGEE in the Schweppe class is generally prohibitive because the full multivariate distribution is needed in order to estimate the regression parameters consistently.

The robustness and efficiency of the REGEE estimators, which is determined by the weights, was investigated in Chapters 4 and 5. In section 4.3.1, an example of medical practice data in which the clusters sizes varied widely, demonstrated the effectiveness of Mallows cluster downweighting REGEE in providing resistant fits. In particular, Mallows REGEE downweighted the cluster with the largest leverage and influence, providing a compromise to the GEE fit based on all the data, and the GEE fit based on the data without the influential cluster. In section 4.3.2, Mallows and Schweppe observation downweighting were applied to a subset of the medical practice data. In this illustration, REGEE served to confirm that there were no disproportionately influential observations and that, in this regard, the GEE fit was adequate. The robustness of REGEE was demonstrated in Chapter 5 by simulating data from a contaminated mixture model for binary responses (Copas, 1988). For a model with 3% contamination, REGEE had a finite sample ($K = 50$) bias that was 40% to 80% smaller than the bias in GEE. For a model with 5% contamination the bias of REGEE was between 15% and 30% smaller than GEE, depending on the design. For a model without contamination, however, the finite sample bias of REGEE was larger than GEE, even though both procedures are asymptotically unbiased.

The efficiency of REGEE with respect to GEE was investigated both asymptotically through analytical means and for small sample sizes through simulations from a model without contamination. General patterns in the asymptotic relative effiiciency of REGEE to GEE were discussed in section 5.3. For correlated binary data, the asymptotic relative efficiency was evaluated in section 5.3.3 for specific designs and the finite sample efficiency of REGEE with respect to GEE was explored in section 5.4 through simulations. Generally speaking, REGEE was found to be less efficient than GEE, but for many practical situations its efficiency is high. Suggestions were made for specification of the tuning constant for Schweppe observation downweighting REGEE, and Mallows downweighting REGEE, based on balancing the desire for robustness and high efficiency.

6.2 Future Research

This work represents an initial effort at addressing the problem of the influence of observations and clusters in GEE. More understanding is needed in the interpretation of the deletion diagnostics presented in Chapter 3 in relation to within cluster correlation and cluster size. The problem of the influence of observations on the naive and robust standard errors of GEE has not been addressed. An extension of *COVRATIO* of Belsley et al. (1980), for example, would estimate the effect of deletion on the covariance matrix in GEE. Also, further extensions might involve the influence on second moments, such as odds ratios or correlations, in the GEE2 procedure of Liang et al. (1992). For GEE, however, one is not usually concerned with the estimation of the nuisance correlation, as long it is consistently estimated. Finally, there are examples of analyses in GEE in which very different regression parameter estimates are obtained under different values of the working correlation. Whether this problem may be related to the influence of a few observations could be explored.

The REGEE methodology described in Chapter 4 requires further development. Limitations of the application of the theory have been addressed in section 4.2.1. For the Schweppe observation downweighting class, one may apply the weights not only to the Pearson residual, but including the leverage as well, as in Pregibon (1982). The theory would apply to the studentized form of the Pearson residual, (2.28), if the part of the weight depending on the correlation is separable from the part of the weight depending on the response as detailed in Appendix 2. In particular, this requirement is met for the weight function applied in section 4.3. REGEE has yet to be applied using other weight functions. A modification of the assumption (*iv*) of Theorem 3 in Chapter 4 might allow the use of non-smooth weight functions such as Huber's function (2.16).

Primary focus in this work was placed on the case of correlated binary responses. The application of REGEE, especially Schweppe observation downweighting, to other types of outcomes, such as Poisson, Gamma and Normal needs development. The robustness and efficiency of REGEE with respect to GEE as studied in Chapter 5, could be studied, beyond the four designs, two values of correlation, and two values of the tuning constant that were considered in this work. A possible extension of the robust GEE idea is to quadratic estimating functions that jointly estimate first and second moment parameters, as in Liang et al. (1992). These might be called Resistant Second Order Generalized Estimating Equations or REGEE2.

APPENDIX 1

Proof of Theorem 1

The proof of THEOREM 1 is obtained by establishing three results for matrices which generalize the results given in the Appendix of Christensen, Pearson, and Johnson (1992a). Their results are generalized here to consider models with a general link function and to consider influence of multiple observations.

Lemma 3.1

$$X_{[m]}^T V_{[m]}^{-1} X_{[m]} = X^T W X - \tilde{X}_m^T W_m \tilde{X}_m.$$

Proof.

$$X^T W X = (X_m^T, X_{[m]}^T)\, V^{-1} \begin{pmatrix} X_m \\ X_{[m]} \end{pmatrix}.$$

We use the following result on the inverse of a partitioned matrix. Let

$$V = \begin{bmatrix} V_m & V_{m[m]} \\ V_{[m]m} & V_{[m]} \end{bmatrix},$$

then

$$V^{-1} = \begin{bmatrix} W_m & -W_m V_{m[m]} V_{[m]}^{-1} \\ -V_{[m]}^{-1} V_{[m]m} W_m & V_{[m]}^{-1} + V_{[m]}^{-1} V_{[m]m} W_m V_{m[m]} V_{[m]}^{-1} \end{bmatrix}$$

where

$$W_m^{-1} = V_m - V_{m[m]} V_{[m]}^{-1} V_{[m]m}.$$

$$\begin{aligned} \text{Thus, } X^T W X = {} & X_m^T W_m X_m - X_m^T W_m V_{m[m]} V_{[m]}^{-1} X_{[m]} - X_{[m]}^T V_{[m]}^{-1} V_{[m]m} W_m X_m + \\ & X_{[m]}^T V_{[m]}^{-1} X_{[m]} + X_{[m]}^T V_{[m]}^{-1} V_{[m]m} W_m V_{m[m]} V_{[m]}^{-1} X_{[m]} \\ = {} & X_{[m]}^T V_{[m]}^{-1} X_{[m]} + (X_m - V_{m[m]} V_{[m]}^{-1} X_{[m]})^T W_m (X_m - V_{m[m]} V_{[m]}^{-1} X_{[m]}) \\ = {} & X_{[m]}^T V_{[m]}^{-1} X_{[m]} + \tilde{X}_m^T W_m \tilde{X}_m \end{aligned}$$

Lemma 3.2

$$(X_{[m]}^T V_{[m]}^{-1} X_{[m]})^{-1} = (X^T W X)^{-1} + (X^T W X)^{-1} \tilde{X}_m^T (W_m^{-1} - \tilde{Q}_m)^{-1} \tilde{X}_m (X^T W X)^{-1}$$

where

$$\tilde{Q}_m = \tilde{X}_m(X^TWX)^{-1}\tilde{X}_m^T$$

Proof.

$$(X_{[m]}^T V_{[m]}^{-1} X_{[m]})^{-1} = (X^TWX - \tilde{X}_m^T W_m \tilde{X}_m)^{-1}$$

$$= (X^TWX)^{-1} + (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{X}_m(X^TWX)^{-1}\tilde{X}_m^T)^{-1}\tilde{X}_m(X^TWX)^{-1}$$

applying result 2.1 above and (17) in chapter 10 of Searle (1982) Matrix Algebra Useful for Statistics.

Lemma 3.3

$$X_{[m]}^T V_{[m]}^{-1} Z_{[m]} = X^TWZ - \tilde{X}_m^T W_m \tilde{Z}_m.$$

Proof. The proof is similar to that of Result 2.1.

$$X^TWZ = (X_m^T, X_{[m]}^T)V^{-1}\begin{pmatrix} Z_m \\ Z_{[m]} \end{pmatrix}$$

$$= X_m^T W_m Z_m - X_m^T W_m V_{m[m]} V_{[m]}^{-1} Z_{[m]} - X_{[m]}^T V_{[m]}^{-1} V_{[m]m} W_m Z_m$$

$$+ X_{[m]}^T V_{[m]}^{-1} Z_{[m]} + X_{[m]}^T V_{[m]}^{-1} V_{[m]m} W_m V_{m[m]} V_{[m]}^{-1} Z_{[m]}$$

$$= X_{[m]}^T V_{[m]}^{-1} Z_{[m]} + (X_m - V_{m[m]} V_{[m]}^{-1} X_{[m]})^T W_m (Z_m - V_{m[m]} V_{[m]}^{-1} Z_{[m]})$$

$$= X_{[m]}^T V_{[m]}^{-1} Z_{[m]} + \tilde{X}_m^T W_m \tilde{Z}_m.$$

Proof of Theorem 1.

$$\hat{\beta}_{[m]} \approx (X_{[m]}^T V_{[m]}^{-1} X_{[m]})^{-1} X_{[m]}^T V_{[m]}^{-1} Z_{[m]}$$

$$= [(X^TWX)^{-1} + (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{X}_m(X^TWX)^{-1}](X^TWZ - \tilde{X}_m^T W_m \tilde{Z}_m)$$

$$= \hat{\beta} + (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{X}_m\hat{\beta} - (X^TWX)^{-1}\tilde{X}_m^T W_m \tilde{Z}_m$$

$$- (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{Q}_m W_m \tilde{Z}_m$$

$$= \hat{\beta} + (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{X}_m\hat{\beta}$$

$$- (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}(W_m^{-1} - \tilde{Q}_m)W_m \tilde{Z}_m$$

$$- (X^TWX)^{-1}\tilde{X}_m^T(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{Q}_m W_m \tilde{Z}_m$$

$$= \widehat{\beta} + (X^T W X)^{-1} \widetilde{X}_m^T (W_m^{-1} - \widetilde{Q}_m)^{-1} \widetilde{X}_m \widehat{\beta} - (X^T W X)^{-1} \widetilde{X}_m^T (W_m^{-1} - \widetilde{Q}_m)^{-1} \widetilde{Z}_m$$

$$= \widehat{\beta} - (X^T W X)^{-1} \widetilde{X}_m^T (W_m^{-1} - \widetilde{Q}_m)^{-1} (\widetilde{Z}_m - \widetilde{X}_m \widehat{\beta})$$

$$= \widehat{\beta} - (X^T W X)^{-1} \widetilde{X}_m^T (W_m^{-1} - \widetilde{Q}_m)^{-1} \widetilde{E}_m.$$

Finally, note that

$$\begin{aligned}
\widetilde{E}_m &= \widetilde{Z}_m - \widetilde{X}_m \widehat{\beta} \\
&= (Z_m - V_{m[m]} V_{[m]}^{-1} Z_{[m]}) - (X_m - V_{m[m]} V_{[m]}^{-1} X_{[m]}) \widehat{\beta} \\
&= \{X_m \widehat{\beta} + D_m (y_m - \widehat{\mu}_m)\} - V_{m[m]} V_{[m]}^{-1} \{X_{[m]} \widehat{\beta} + D_{[m]} (y_{[m]} - \widehat{\mu}_{[m]})\} \\
&\qquad - X_m \widehat{\beta} + V_{m[m]} V_{[m]}^{-1} X_{[m]} \widehat{\beta} \\
&= D_m (y_m - \widehat{\mu}_m) - V_{m[m]} V_{[m]}^{-1} D_{[m]} (y_{[m]} - \widehat{\mu}_{[m]}) \\
&\qquad\qquad\qquad\qquad = E_m - V_{m[m]} V_{[m]}^{-1} E_{[m]}.
\end{aligned}$$

APPENDIX 2

Proof of Theorem 3: Schweppe observation downweighting class

We establish consistency and asymptotic normality of the regression parameter estimates from the Schweppe class of Resistant Generalized Estimating Equations (ReGEE),

$$\sum_{i=1}^{k} D_i'(\beta)V_i^{-1}(\alpha,\ \beta)\{W_{ki}(Y_i, \alpha,\ \beta)(Y_i - \mu_i(\beta)) - c_i\},$$

under the special condition that $W_{ki}(Y_i, \alpha,\ \beta) = W_{i1}(Y_i, \beta)W_{i2}(\alpha)$. In other words, the REGEE weight for each observation partitions into a part that depends on the data but not α, and a second part that depends on α, but not the data. Under this constraint, the Schweppe class ReGEE have individual contributions $U_{ki} = D_i'V_i^{-1}(\psi_{ki} - c_i)$ that are products of two terms: the first involving α but not the data, and the second independent of α and with expectation zero. This can be seen by writing,

$$U_{ki} = D_i'(\beta)V_i^{-1}(\alpha, \beta)\{W_{ki1}(Y_i, \beta)W_{ki2}(\alpha)(Y_i - \mu_i(\beta)) - W_{ki2}(\alpha)c_{i1}\}$$

$$= D_i'(\beta)W_{ki2}^{*\prime}\{W_{ki1}(Y_i, \beta)(Y_i - \mu_i(\beta)) - c_{i1}(\beta)\}$$

where $W_{ki2}^{*}(\alpha, \beta) = W_{ki2}(\alpha)V_i^{-1}(\alpha, \beta)$ and $c_{i1}(\beta) = E\Big\{W_{ki1}(Y_i, \beta)(Y_i - \mu_i(\beta))\Big\}$. Note that c_{i1} will not depend on α in the Schweppe observation downweighting class. For the Schweppe cluster downweighting class, however, it may depend on α, in which case the proof does not apply.

Additionally, the following regularity conditions are required for the proof:

(v) $|\partial\,\widehat{\alpha}\,(\beta, \phi)/\partial\phi| \le H_1(Y, \beta)$ *which is* $O_p(1)$, *for all* β *and* ϕ;

(vi) $|\partial\,\widehat{\alpha}\{\beta, \widehat{\phi}(\beta)\}/\partial\beta| \le H_2(Y, \beta)$ *which is* $O_p(1)$, *for all* β and $\widehat{\phi}$;

(vii) $|\partial\widehat{\phi}(\beta)/\partial\beta| \le H_3(Y,\beta)$ *which is* $O_p(1)$, *for all* β;

The proof follows along the lines of the proof of Theorem 2, in the Appendix of Liang and Zeger (1986).

Write $\alpha^*(\beta) = \widehat{\alpha}\{\beta, \widehat{\phi}(\beta)\}$, and a Taylor expansion about β is

$$U\{\widehat{\beta},\ \alpha^*(\widehat{\beta})\} = 0 = U\{\beta,\ \alpha^*(\beta)\} + \tfrac{\delta}{\delta\beta}U\{\beta,\ \alpha^*(\beta)\}\Big|_{\beta=\widetilde{\beta}}(\beta - \widehat{\beta})$$

for some $\widetilde{\beta} = \lambda\beta + (1 - \lambda)\widehat{\beta}$, $\lambda \in (0,\ 1)$.

It follows that

$$k^{1/2}(\widehat{\beta}-\beta)=\frac{\sum_{i=1}^{k}U_i\{\beta,\,\alpha^*(\beta)\}/k^{1/2}}{\frac{1}{k}\sum_{i=1}^{k}\frac{\delta}{\delta\beta}U_i\{\beta,\alpha^*(\beta)\}\Big|_{\beta=\widetilde{\beta}}} \tag{A.2.1}$$

where

$$\tfrac{\delta}{\delta\beta}U_i\{\beta,\,\alpha^*(\beta)\}=\tfrac{\partial}{\partial\beta}U_i\{\beta,\,\alpha^*\}+\tfrac{\partial}{\partial\alpha^*}U_i\{\beta,\,\alpha^*(\beta)\}\tfrac{\delta\alpha^*(\beta)}{\delta\beta}=A_i+B_iC.$$

The first part of the proof is to show asymptotic normality of the numerator of (A.2.1). A Taylor expansion about α gives

$$\begin{aligned}\sum_{i=1}^{k}U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2}&=\sum_{i=1}^{k}U_i(\beta,\ \alpha)/k^{1/2}+\tfrac{1}{k}\sum_{i=1}^{k}\tfrac{\partial}{\partial\alpha}U_i(\beta,\,\alpha)\Big|_{\alpha=\widetilde{\alpha}}k^{1/2}(\alpha-\alpha^*)\\&=A^*+B^*C^*\end{aligned}$$

for some $\widetilde{\alpha}=\lambda\alpha+(1-\lambda)\widehat{\alpha},\ \lambda\in(0,\,1)$.

Consider B^*. Let $S_i(Y_i,\beta)=W_{ki1}(Y_i,\beta)(Y_i-\mu_i(\beta))-c_{i1}$, and write

$$U_i(\beta,\,\alpha)=D_i'(\beta)W_{ki2}^{*\prime}(\alpha,\,\beta)S_i(Y_i,\beta),$$

and note that

$$E\{\frac{\partial}{\partial\alpha}U_i(\beta,\,\alpha)\Big|_{\alpha=\widetilde{\alpha}}\}=0, \tag{A.2.2}$$

because $S_i(Y_i,\beta)$ is independent of α and has expectation zero.

Let Z_{imn} be the element from the m-th row and the n-th column of the matrix $\frac{\partial}{\partial\alpha}U_i(\beta,\alpha)\Big|_{\alpha=\widetilde{\alpha}}$, and assume that $k^{-1-\delta}\sum_{i=1}^{k}E|Z_{imn}|^{1+\delta}\to 0$ as $k\to\infty$ for all m, n (i.e., the Markov Condition). Then $B^*\xrightarrow{p}0$ by the Markov Weak Law of Large Numbers. Equivalently $B^*=o_p(1)$.

Next consider C^*. Write

$$C^*=k^{1/2}[\alpha-\widehat{\alpha}(\beta,\,\phi)+\widehat{\alpha}(\beta,\,\phi)-\widehat{\alpha}\{\beta,\,\widehat{\phi}(\beta)\}]. \tag{A.2.3}$$

A Taylor expansion about ϕ gives

$$\widehat{\alpha}\{\beta,\,\widehat{\phi}(\beta)\}=\widehat{\alpha}(\beta,\,\phi)+\frac{\partial\widehat{\alpha}(\beta,\,\phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}(\phi-\widehat{\phi}) \tag{A.2.4}$$

where $\widetilde{\phi}=(\phi,\,\widehat{\phi})$.

Substituting (A.2.4) into (A.2.3) yields

$$C^* = k^{1/2}\Big\{\alpha - \widehat{\alpha}(\beta,\, \phi) + \frac{\partial\widehat{\alpha}(\beta,\phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}(\phi - \widehat{\phi}(\beta))\Big\}.$$

By condition (i), $k^{1/2}(\alpha - \widehat{\alpha}(\beta,\, \phi)) = O_p(1)$;
by condition (ii), $k^{1/2}(\phi - \widehat{\phi}(\beta)) = O_p(1)$;
and condition (v) implies that $\frac{\partial\widehat{\alpha}(\beta,\phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}$ is bounded in probability. Thus $C^* = O_p(1)$.

Consequently,

$$\sum_{i=1}^{k} U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2} = A^* + O_p(1). \tag{A.2.5}$$

Now observe that $E(A^*) = 0$, and

$$\begin{aligned} Var(A^*) &= \tfrac{1}{k}\sum_{i=1}^{k} Var[\, D_i'(\beta)W_{ki2}^{*\prime}(\alpha,\, \beta)S_i(Y_i,\beta)] \\ &= \tfrac{1}{k}\sum_{i=1}^{k} D_i'W_{ki2}^{*\prime}Var(\psi_{i1})W_{ki2}^{*}D_i \\ &= \tfrac{1}{k}\sum_{i=1}^{k} D_i'V_i^{-1}Var(\psi_i)V_i^{-1}D_i \end{aligned}$$

where $\psi_{i1} = W_{ki1}(Y_i,\beta)(Y_i - \mu_i(\beta))$ and $\psi_i = W_{ki}(Y_i,\alpha,\, \beta)(Y_i - \mu_i(\beta))$.

By the Central Limit Theorem for a triangular array of univariate random variables (Theorem 3.3.5 of Singer and Sen; 1993) and the Cramér-Wold device, (Theorem 3.2.4 of Singer and Sen, 1993),

$$A^* \sim MVN(0,\ \lim_{k\to\infty} Var(A^*)). \tag{A.2.6}$$

The details follow. For some arbitrary vector λ, let $\lambda' A^* = \sum_{i=1}^{k}\lambda' U_{ki}/k^{1/2}$ and $s_k^2 = Var(\lambda' A^*)$.

For the i-th cluster define $X_{ki} = \frac{\lambda' U_{ki}/k^{1/2}}{s_k}$.

Then $\{X_{ki},\ 1 \le i \le k\}$ is a triangular array of random variables, such that for each k, $\{X_{ki},\ 1 \le i \le k\}$ are independent. [X_{ki} should not be confused with X_i, the covariate matrix for the i-th cluster, rather the notation is chosen to be consistent with the notation in Sen and Singer (1993)].

Now, in order to show

$$\sum_{i=1}^{k} X_{ki} = \frac{\lambda' A^*}{s_k} \xrightarrow{\mathcal{D}} N(0,\,1), \tag{A.2.7}$$

conditions A and B of theorem 3.3.5 of Sen and Singer (p.118) are assumed. First, note that condition A holds:

$$\sum_{i=1}^{k} \mathrm{P}\{|X_{ki}| > \epsilon\} = \sum_{i=1}^{k} \mathrm{P}\{|\lambda' U_{ki}/k^{1/2}| > \epsilon\, s_k\} \to 0$$

as $k \to \infty$ for $\epsilon > 0$.

Second, we require condition B,

$$\sum_{i=1}^{k} \Big\{ EX_{ki}^2 I\{|X_{ki}| \le \epsilon\} - [EX_{ki} I\{|X_{ki}| \le \epsilon\}]^2 \Big\} \to 1$$

Condition B is implied by the fact $\sum_{i=1}^{k}\{EX_{ki}^2 - [EX_{ki}]^2\} = Var(\sum_{i=1}^{k} X_{ki}) = 1$ and by assuming

$$\sum_{i=1}^{k} [EX_{ki}^2 I\{|X_{ki}| > \epsilon\}] \to 0, \qquad \text{and} \qquad \sum_{i=1}^{k} [EX_{ki} I\{|X_{ki}| > \epsilon\}]^2 \to 0.$$

The first is essentially the Lindeberg-Feller condition.

Conditions A and B of Theorem 3.3.5 of Singer and Sen, impl that $\{X_{ki}\}$ is an infinitesimal system of random variables, and (A.2.7) holds. By the Cramér-Wold device, (A.2.7) implies (A.2.6). From an application of Slutsky's Theorem to (A.2.5) and (A.2.6) we obtain,

$$\sum_{i=1}^{k} U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2} \sim MVN(0, \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} D_i' W_{ki}^{*\prime} Var(Y_i) W_{ki}^* D_i) \tag{A.2.8}$$

This completes the first part of the proof.

The second and final part of the proof considers the denominator on the right hand side of (A.2.1).

Consider B_i. Because $S_i(Y_i, \beta)$ is independent of α^*, like (A.2.2), we obtain $E(B_i) = 0$.

Thus $\sum_{i=1}^{k} B_i/k \xrightarrow{p} 0$.

Next consider C. Write

$$C = \frac{\delta\alpha^*(\beta)}{\delta\beta} = \frac{\partial\alpha^*(\beta,\widehat{\phi})}{\partial\beta} + \frac{\partial\alpha^*(\beta,\widehat{\phi})}{\partial\widehat{\phi}}\frac{\partial\widehat{\phi}(\beta)}{\partial\beta}.$$

Each part of the right hand side is bounded in probability under conditions (iv), (ii) and (iii) jointly, and (v). Thus $C = O_p(1)$.

Finally, consider $A_i = \frac{\partial}{\partial\beta}U_i\{\beta,\, \alpha^*\} = \frac{\partial}{\partial\beta}\{D_i'(\beta)V_i^{-1}(\alpha^*,\, \beta)(\psi_i - c_i)\}$

First, it is shown that,

$$\mathrm{E}[\partial\mathrm{U}(\beta,\, \alpha)/\partial\beta] = \sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i$$

where $\Gamma_i = \mathrm{E}\dot{\psi}_{ki} - \dot{c}_i$, $\dot{\psi}_{ki} = \frac{\partial}{\partial\mu_i}\psi_i(\mu_i)$ and $\dot{c}_i = \frac{\partial}{\partial\mu_i}c_i$.

By the chain rule, $\partial\mathrm{U}(\beta,\, \alpha)/\partial\beta = \frac{\partial}{\partial\beta}\{D_i'V_i^{-1}\}\left[\psi_{ki} - c_i\right] + \{D_i'V_i^{-1}\}\frac{\partial}{\partial\beta}\left[\psi_{ki} - c_i\right]$.

The first part has expectation zero by condition (iii), and in the second part, under condition (iv), apply,

$$\frac{\partial}{\partial\beta}\left[\psi_{ki} - c_i\right] = \frac{\partial}{\partial\mu}\left[\psi_{ki} - c_i\right]\frac{\partial\mu_i}{\partial\beta} = \left[\dot{\psi}_{ki} - \dot{c}_i\right]D_i.$$

Then by the Markow Weak Law of Large Numbers,

$$\frac{1}{k}\,\partial\mathrm{U}(\beta,\, \alpha)/\partial\beta \xrightarrow{p} \Gamma.$$

where $\Gamma = \lim_{k \to \infty} -\frac{1}{k}\sum_{i=1}^{k} D_i' V_i^{-1} \Gamma_i D_i$.

Under certain regularity conditions (continuity, compact boundedness):

$$\frac{1}{k}\left|\left|\sum_{i=1}^{k}\frac{\partial}{\partial\beta}U_i(\beta,\, \alpha^*) - \sum_{i=1}^{k}\frac{\partial}{\partial\beta}U_i(\beta,\, \alpha)\right|\right| \xrightarrow{p} 0$$

It follows that

$$\frac{1}{k}\sum_{i=1}^{k} A_i \xrightarrow{p} \Gamma. \tag{A.2.9}$$

Given that $\sum_{i=1}^{k} B_i/k$ is $o_p(1)$ and $C = O_p(1)$, it follows that

$$J(\beta)/k = \tfrac{1}{k}\sum_{i=1}^{k} A_i + o_p(1),$$

where $J(\beta) = \sum_{i=1}^{k} \frac{\delta}{\delta\beta} U_i\{\beta,\ \alpha^*(\beta)\}$. From (A.2.9) it follows that

$$J(\beta)/k \xrightarrow{p} \Gamma.$$

To complete the proof we need to show that

$$J(\tilde{\beta})/k \xrightarrow{p} \Gamma \tag{A.2.10}$$

where $J(\tilde{\beta})/k$ is the denominator in (A.2.1).

Assume that $J(\tilde{\beta})$ is continuous in a neighborhood of β. Furthermore, assume boundedness over compact neighborhood, i.e.,

$$P\Big(\sup_{\|\tilde{\beta} - \beta\| < hk^{-1/2}} \Big\| J(\tilde{\beta}) - J(\beta) \Big\| > \delta \Big) \xrightarrow[k \to \infty]{} 0$$

for all $h > 0$ such that δ>0 exists. These regularity conditions imply

$$\tfrac{1}{k}\Big\| J(\tilde{\beta}) - J(\beta) \Big\| \xrightarrow{p} 0$$

which implies (A.2.10).

Finally applying Slutsky's Theorem to (A.2.1), (A.2.8) and (A.2.10) we obtain the result.□

APPENDIX 3

Proof of Theorem 3: Mallows class

We establish consistency and asymptotic normality of the regression parameter estimates from the Mallows class of Resistant Generalized Estimating Equations (ReGEE),

$$\sum_{i=1}^{k} D_i'(\beta)V_i^{-1}(\alpha,\,\beta)W_{ki}(\alpha,\,\beta)(Y_i-\mu_i(\beta)).$$

The proof follows along the lines of the proof of Theorem 2, in the Appendix of Liang and Zeger (1986). As in GEE, the Mallows class ReGEE have individual contributions $U_{ki} = D_i'V_i^{-1}W_{ki}(Y_i-\mu_i)$ that are products of two terms: the first involving α but not the data, and the second independent of α and with expectation zero. Additionally, the following regularity conditions are required for the proof:

(v) $\quad |\partial\,\widehat{\alpha}\,(\beta,\phi)/\partial\phi| \;\le H_1(Y,\beta)\,which\,is\,O_p(1),\;for\,all\,\beta\,and\,\phi;$

(vi) $\quad |\partial\,\widehat{\alpha}\{\beta,\widehat{\phi}(\beta)\}/\partial\beta| \;\le H_2(Y,\beta)\,which\,is\,O_p(1),\;for\,all\,\beta$ and $\widehat{\phi}$;

(vii) $\quad |\partial\widehat{\phi}(\beta)/\partial\beta| \le H_3(Y,\beta)\,which\,is\,O_p(1),\;for\,all\,\beta;$

Conditions (iii) and (iv) pertain to the Schweppe proof (Appendix 3) only.

Write $\alpha^*(\beta) = \widehat{\alpha}\{\beta,\widehat{\phi}(\beta)\}$, and a Taylor expansion about β is

$$U\{\widehat{\beta},\,\alpha^*(\widehat{\beta})\} = 0 = U\{\beta,\,\alpha^*(\beta)\} + \tfrac{\delta}{\delta\beta}U\{\beta,\,\alpha^*(\beta)\}\Big|_{\beta=\widetilde{\beta}}(\beta-\widehat{\beta})$$

for some $\widetilde{\beta} = \lambda\beta + (1-\lambda)\widehat{\beta},\ \lambda\in(0,\,1)$.

It follows that

$$k^{1/2}(\widehat{\beta}-\beta) = \frac{\sum_{i=1}^{k} U_i\{\beta,\,\alpha^*(\beta)\}/k^{1/2}}{\frac{1}{k}\sum_{i=1}^{k}\frac{\delta}{\delta\beta}U_i\{\beta,\alpha^*(\beta)\}\Big|_{\beta=\widetilde{\beta}}} \tag{A.3.1}$$

where

$$\tfrac{\delta}{\delta\beta}U_i\{\beta,\,\alpha^*(\beta)\} = \tfrac{\partial}{\partial\beta}U_i\{\beta,\,\alpha^*\} + \tfrac{\partial}{\partial\alpha^*}U_i\{\beta,\,\alpha^*(\beta)\}\tfrac{\delta\alpha^*(\beta)}{\delta\beta} = A_i + B_iC.$$

The first step of the proof is to show asymptotic normality of the numerator of (A.3.1). A Taylor expansion about α gives

$$\sum_{i=1}^{k} U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2} = \sum_{i=1}^{k} U_i(\beta,\ \alpha)/k^{1/2} + \tfrac{1}{k}\sum_{i=1}^{k} \tfrac{\partial}{\partial\alpha} U_i(\beta,\ \alpha)\Big|_{\alpha=\widetilde{\alpha}} k^{1/2}(\alpha - \alpha^*)$$
$$= A^* + B^* C^*$$

for some $\widetilde{\alpha} = \lambda\alpha + (1-\lambda)\widehat{\alpha}$, $\lambda \in (0,\ 1)$.

Consider B^*. Let $S_i = Y_i - \mu_i$, and $W_{ki}^* = W_{ki} V_i^{-1}$. Then write

$$U_i(\beta,\ \alpha) = D_i'(\beta) W_{ki}^{*\prime}(\alpha,\ \beta) S_i(\beta),$$

and note that

$$E\{\frac{\partial}{\partial\alpha} U_i(\beta,\ \alpha)\Big|_{\alpha=\widetilde{\alpha}}\} = 0, \tag{A.3.2}$$

because $S_i(\beta)$ is independent of α and has expectation zero.

Let Z_{imn} be the element from the m-th row and the n-th column of the matrix $\frac{\partial}{\partial\alpha} U_i(\beta,\alpha)\Big|_{\alpha=\widetilde{\alpha}}$, and assume that $k^{-1-\delta}\sum_{i=1}^{k} E|Z_{imn}|^{1+\delta} \to 0$ as $k \to \infty$ for all m, n (i.e., the Markov Condition). Then $B^* \xrightarrow{p} 0$ by the Markov Weak Law of Large Numbers. Equivalently $B^* = o_p(1)$.

Next consider C^*. Write

$$C^* = k^{1/2}[\alpha - \widehat{\alpha}(\beta,\ \phi) + \widehat{\alpha}(\beta,\ \phi) - \widehat{\alpha}\{\beta,\ \widehat{\phi}(\beta)\}]. \tag{A.3.3}$$

A Taylor expansion about ϕ gives

$$\widehat{\alpha}\{\beta,\ \widehat{\phi}(\beta)\} = \widehat{\alpha}(\beta,\ \phi) + \frac{\partial\widehat{\alpha}(\beta,\ \phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}(\phi - \widehat{\phi}) \tag{A.3.4}$$

where $\widetilde{\phi} = (\phi,\ \widehat{\phi})$.

Substituting (A.3.4) into (A.3.3) yields

$$C^* = k^{1/2}\Big\{\alpha - \widehat{\alpha}(\beta,\ \phi) + \tfrac{\partial\widehat{\alpha}(\beta,\phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}(\phi - \widehat{\phi}(\beta))\Big\}.$$

By condition (i), $k^{1/2}(\alpha - \widehat{\alpha}(\beta,\ \phi)) = O_p(1)$;
by condition (ii), $k^{1/2}(\phi - \widehat{\phi}(\beta)) = O_p(1)$;
and condition (v) implies that $\frac{\partial\widehat{\alpha}(\beta,\phi)}{\partial\phi}\Big|_{\phi=\widetilde{\phi}}$ is bounded in probability. Thus $C^* = O_p(1)$.

Consequently,

$$\sum_{i=1}^{k} U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2} = A^* + O_p(1). \tag{A.3.5}$$

Now observe that $E(A^*) = 0$, and

$$Var(A^*) = \tfrac{1}{k}\sum_{i=1}^{k} Var[\ D_i'(\beta)W_{ki}^{*\prime}(\alpha,\ \beta)S_i(\beta)]$$

$$= \tfrac{1}{k}\sum_{i=1}^{k} D_i' W_{ki}^{*\prime} Var(Y_i) W_{ki}^{*} D_i.$$

By the Central Limit Theorem for a triangular array of univariate random variables (Theorem 3.3.5 of Singer and Sen; 1993) and the Cramér-Wold device, (Theorem 3.2.4 of Singer and Sen, 1993),

$$A^* \sim MVN(0,\ \lim_{k \to \infty} Var(A^*)). \tag{A.3.6}$$

The details follow. For some arbitrary vector λ, let $\lambda' A^* = \sum_{i=1}^{k} \lambda' U_{ki}/k^{1/2}$ and $s_k^2 = Var(\lambda' A^*)$.

For the i-th cluster define $X_{ki} = \frac{\lambda' U_{ki}/k^{1/2}}{s_k}$.

Then $\{X_{ki},\ 1 \le i \le k\}$ is a triangular array of random variables, such that for each k, $\{X_{ki},\ 1 \le i \le k\}$ are independent. [X_{ki} should not be confused with X_i, the covariate matrix for the i-th cluster, rather the notation is chosen to be consistent with the notation in Sen and Singer (1993)].

Now, in order to show

$$\sum_{i=1}^{k} X_{ki} = \frac{\lambda' A^*}{s_k} \xrightarrow{\mathcal{D}} N(0,\ 1), \tag{A.3.7}$$

conditions A and B of theorem 3.3.5 of Sen and Singer (p.118) are assumed. First, note that condition A holds:

$$\sum_{i=1}^{k} \mathrm{P}\{|X_{ki}| > \epsilon\} = \sum_{i=1}^{k} \mathrm{P}\{|\lambda' U_{ki}/k^{1/2}| > \epsilon\ s_k\} \to 0$$

as $k \to \infty$ for $\epsilon > 0$.

Second, we require condition B,

$$\sum_{i=1}^{k}\left\{EX_{ki}^2 I\{|X_{ki}| \le \epsilon\} - [EX_{ki} I\{|X_{ki}| \le \epsilon\}]^2\right\} \to 1$$

Condition B is implied by the fact $\sum_{i=1}^{k}\{EX_{ki}^2 - [EX_{ki}]^2\} = Var(\sum_{i=1}^{k} X_{ki}) = 1$ and by assuming

$$\sum_{i=1}^{k}[EX_{ki}^2 I\{|X_{ki}| > \epsilon\}] \to 0, \quad \text{and} \quad \sum_{i=1}^{k}[EX_{ki} I\{|X_{ki}| > \epsilon\}]^2 \to 0.$$

The first is essentially the Lindeberg-Feller condition.

Conditions A and B of Theorem 3.3.5 of Singer and Sen, impl that $\{X_{ki}\}$ is an infinitesimal system of random variables, and (A.3.7) holds. By the Cramér-Wold device, (A.3.7) implies (A.3.6). From an application of Slutsky's Theorem to (A.3.5) and (A.3.6) we obtain,

$$\sum_{i=1}^{k} U_i\{\beta,\ \alpha^*(\beta)\}/k^{1/2} \sim MVN(0, \lim_{k \to \infty} \frac{1}{k}\sum_{i=1}^{k} D_i' W_{ki}^{*\prime} Var(Y_i) W_{ki}^* D_i) \quad \text{(A.3.8)}$$

This completes the first part of the proof.

The second and final part of the proof considers the denominator on the right hand side of (A.3.1).

Consider B_i. Because $S_i(\beta)$ is independent of α^*, like (A.3.2), we obtain $E(B_i) = 0$. Thus $\sum_{i=1}^{k} B_i/k \xrightarrow{p} 0$.

Next consider C. Write

$$C = \frac{\delta\alpha^*(\beta)}{\delta\beta} = \frac{\partial\alpha^*(\beta,\widehat{\phi})}{\partial\beta} + \frac{\partial\alpha^*(\beta,\widehat{\phi})}{\partial\widehat{\phi}}\frac{\partial\widehat{\phi}(\beta)}{\partial\beta}.$$

Each part of the right hand side is bounded in probability under conditions (*vi*), (*ii*) and (*v*) jointly, and (*vii*). Thus $C = O_p(1)$.

Finally, consider $A_i = \frac{\partial}{\partial\beta} U_i\{\beta,\ \alpha^*\}$

$$= \frac{\partial}{\partial\beta}\{D_i'(\beta) W_{ki}^{*\prime}(\alpha^*,\ \beta) S_i(\beta)\}$$

By the chain rule, the fact that $E(S_i) = 0$, and the Markov Weak Law of Large Numbers,

$$\frac{1}{k}\sum_{i=1}^{k}\frac{\partial}{\partial\beta}U_i(\beta,\,\alpha)\xrightarrow{p}\Gamma$$

where $\Gamma = \lim_{k\to\infty} -\frac{1}{k}\sum_{i=1}^{k}D_i'W_{ki}^{*\prime}D_i$.

Under certain regularity conditions (continuity, compact boundedness):

$$\frac{1}{k}\left|\left|\sum_{i=1}^{k}\frac{\partial}{\partial\beta}U_i(\beta,\,\alpha^*)-\sum_{i=1}^{k}\frac{\partial}{\partial\beta}U_i(\beta,\,\alpha)\right|\right|\xrightarrow{p}0$$

It follows that

$$\frac{1}{k}\sum_{i=1}^{k}A_i\xrightarrow{p}\Gamma. \tag{A.3.9}$$

Given that $\sum_{i=1}^{k}B_i/k$ is $o_p(1)$ and $C = O_p(1)$, it follows that

$$J(\beta)/k = \frac{1}{k}\sum_{i=1}^{k}A_i + o_p(1),$$

where $J(\beta) = \sum_{i=1}^{k}\frac{\delta}{\delta\beta}U_i\{\beta,\,\alpha^*(\beta)\}$. From (A.1.9) it follows that

$$J(\beta)/k\xrightarrow{p}\Gamma.$$

To complete the proof we need to show that

$$J(\widetilde{\beta})/k\xrightarrow{p}\Gamma \tag{A.3.10}$$

where $J(\widetilde{\beta})/k$ is the denominator in (A.3.1).

Assume that $J(\widetilde{\beta})$ is continuous in a neighborhood of β. Furthermore, assume boundedness over compact neighborhood, i.e.,

$$P\Big(\sup_{||\widetilde{\beta}-\beta||<hk^{-1/2}}\left|\left|J(\widetilde{\beta})-J(\beta)\right|\right|>\delta\Big)\xrightarrow[k\to\infty]{}0$$

for all $h > 0$ such that δ>0 exists. These regularity conditions imply

$$\frac{1}{k}\left|\left|J(\widetilde{\beta})-J(\beta)\right|\right|\xrightarrow{p}0$$

which implies (A.3.10).

Finally applying Slutsky's Theorem to (A.3.1), (A.3.8) and (A.3.10) we obtain the result.

APPENDIX 4

Verification of the conditions of Theorem 2 for Schweppe observation downweighting for correlated binary responses

Conditions (i) through (iv) of Theorem 2 in Chapter 4 are verified for the application of Schweppe observation downweighting for correlated binary responses.

(i) Observing that $E(r^*_{it}r^*_{it'}) = \rho_{itt'}b_{it}b_{it'}\phi$ it can be shown that $\hat{\alpha}$ in (4.6) is $k^{1/2}$-consistent given β and ϕ by the Kolmogorov Law of Large Numbers (Theorem 2.3.10 of Sen and Singer (1993)). In particular, define

$$\sigma_i^2 = \mathrm{Var}(\sum_{t>t'} r^*_{it}r^*_{it'}).$$

Then $\hat{\alpha}$ is strongly consistent, for bounded random variables, like the bernoulli, because

$$\sum_{k\geq 1} k^{-2}\sigma_k^2 < \sum_{k\geq 1} k^{-2}M = M\sum_{k\geq 1} k^{-2} < \infty$$

if the n_i are bounded. For other random variables, like the Poissoin, additional assumptions about the design are required for the condition to hold.

(ii) Observing that $E(r^{*2}_{it}) = b^2_{it}\phi$ it can be shown that $\hat{\phi}$ in (4.5) is $k^{1/2}$-consistent given β by the Kolmogorov Law of Large Numbers (Theorem 2.3.10 of Sen and Singer (1993)), where the above condition holds for $\sigma_i^2 = \mathrm{Var}(\sum_{t=1}^{n_i} r^{*2}_{it})$.

(iii) $0 \leq b_{it} \leq 1$ and $0 \leq v_{it} \leq 1$ imply that $\mathrm{Var}(\psi_{it}) < \infty$ and $\mathrm{Cov}(\psi_{it},\psi_{it'}) < \infty$.

(iv) The weight function, $w_{it}(r_{it})$, must be chosen such that ψ_i is absolutely continuous in β and $\dot{\psi}$ exists for almost all Y. Observe that

$$\dot{\psi}_{it} = \frac{\partial\psi_{it}}{\partial\mu_{it}} = \frac{\partial w_{it}}{\partial r_{it}}\dot{r}_{it}(Y_{it} - \mu_{it}) - w_{it}.$$

It follows from

$$E||\frac{\partial W_i}{\partial r_i}\dot{r}_i(Y_i - \mu_i) - W_i|| \;\leq\; E||\frac{\partial W_i}{\partial r_i}\dot{r}_i(Y_i - \mu_i)|| + E||W_i||$$

that a sufficient condition for $E||\,\dot{\psi}\,|| < \infty$ is that $\frac{\partial w_{it}}{\partial r_{it}}$ and $\dot{r}_{it}$ are finite, $t = 1, ..., n_i$. This will be true for bounded random variables such as the Bernoulli.

APPENDIX 5

Schweppe cluster downweighting for bivariate binary responses

Consider the situation of two correlated binary outcomes. Further, suppose the clusters are downweighted according to an aggregate measure of the lack of fit of the observations in the cluster given by

$$e_i = \frac{1}{2}\sum_{t=1}^{2} r_{it}^2.$$

To solve the Schweppe REGEE, (4.1), we need to find $c_i = (c_{i1}, c_{i2})'$ from the bivariate distribution of (Y_1, Y_2). For an arbitrary cluster and suppressing the subscript i, define w_{jk} as the weight evaluated at $Y_1 = j$ and $Y_2 = k$. Then the bias is

$$\mathrm{c} = \begin{bmatrix} 1-\mu_1 \\ 1-\mu_2 \end{bmatrix} \pi_{11} w_{11} + \begin{bmatrix} 1-\mu_1 \\ -\mu_2 \end{bmatrix} \pi_{10} w_{10} + \begin{bmatrix} -\mu_1 \\ 1-\mu_2 \end{bmatrix} \pi_{01} w_{01} + \begin{bmatrix} -\mu_1 \\ -\mu_2 \end{bmatrix} \pi_{00} w_{00}.$$

Thus, the bias depends on the correlation, ρ. If ρ is known, Theorem 3 applies. The proof in Appendix 2, however, does not allow the bias to depend on an unknown correlation, and so does not pertain to Schweppe cluster downweighting. It is conjectured, however, that this assumption may be relaxed by considering the bias as a function of ρ, providing a consistent estimate of ρ and modifying Taylor series expansions about ρ in the proof.

Assuming, under the conditions of Theorem 3, that the weight function, w, is smooth and differentiable, and that we can interchange expectation and differentiation with respect to the marginal means, it can be shown that

$$\Gamma_i = \mathrm{E}\begin{bmatrix} \frac{\partial \psi_1}{\partial \mu_1} & \frac{\partial \psi_1}{\partial \mu_2} \\ \frac{\partial \psi_2}{\partial \mu_1} & \frac{\partial \psi_2}{\partial \mu_2} \end{bmatrix} - \begin{bmatrix} \frac{\partial c_1}{\partial \mu_1} & \frac{\partial c_1}{\partial \mu_2} \\ \frac{\partial c_2}{\partial \mu_1} & \frac{\partial c_2}{\partial \mu_2} \end{bmatrix}$$

$$= \begin{bmatrix} 1-\mu_1 \\ 1-\mu_2 \end{bmatrix} [-g_1, -g_2 \quad] w_{11} + \begin{bmatrix} 1-\mu_1 \\ -\mu_2 \end{bmatrix} [-(1-g_1), g_2 \quad] w_{10}$$

$$+ \begin{bmatrix} -\mu_1 \\ 1-\mu_2 \end{bmatrix} [-g_1, -(1-g_2) \quad] w_{01} + \begin{bmatrix} -\mu_1 \\ -\mu_2 \end{bmatrix} [(1-g_1), (1-g_2) \quad] w_{00},$$

where $g_1 = \mu_2 + \frac{(1-2\mu_1)\rho v_2^{1/2}}{2v_1^{1/2}}$ and $g_2 = \mu_1 + \frac{(1-2\mu_2)\rho v_1^{1/2}}{2v_2^{1/2}}$.

The variance estimate of $\hat{\beta}_S$ is given by (4.2) and

$$\mathrm{Var}(\psi) = \begin{bmatrix} 1-\mu_1 \\ 1-\mu_2 \end{bmatrix} [1-\mu_1, 1-\mu_2 \quad] \pi_{11} w_{11}^2 + \begin{bmatrix} 1-\mu_1 \\ -\mu_2 \end{bmatrix} [1-\mu_1, -\mu_2 \quad] \pi_{10} w_{10}^2$$

$$+ \begin{bmatrix} -\mu_1 \\ 1-\mu_2 \end{bmatrix} [-\mu_1, 1-\mu_2 \quad] \pi_{01} w_{01}^2 + \begin{bmatrix} -\mu_1 \\ -\mu_2 \end{bmatrix} [\mu_1, \mu_2 \quad] \pi_{00} w_{00}^2,$$

In practice, however, the variance of $\hat{\beta}_S$ is estimated by (4.3). We can not apply (4.6) because (4.4) does not hold, but a method of moments estimator for ρ can be determined by observing that

$$\mathrm{E}(wr_1 wr_2) = \frac{\mathrm{E}(\psi_1\psi_2)}{v_1^{1/2} v_2^{1/2}} = A\rho + \frac{\mu_1\mu_2}{v_1^{1/2} v_2^{1/2}}(A-B)$$

where $A = (1-\mu_1)(1-\mu_2)w_{11}^2 + (1-\mu_1)\mu_2 w_{10}^2 + \mu_1(1-\mu_2)w_{01}^2 + \mu_1\mu_2 w_{00}^2$
and $B = (1-\mu_1)w_{10}^2 + (1-\mu_2)w_{01}^2 - (1-\mu_1-\mu_2)w_{00}^2$.

Thus,

$$\hat{\rho} = \frac{\sum_{i=1}^{K} [(\psi_{i1}\psi_{i2} - \mu_1\mu_2(A-B)]/[v_1^{1/2} v_2^{1/2}]}{\sum_{i=1}^{K} A_i - p}$$

with p added as a correction factor as in Liang and Zeger (1986).

APPENDIX 6

Proof of non-optimality result in section 5.2

Consider Schweppe observation downweighting for correlated binary responses with logit link and variance function equal to $v_{it} = \mu_{it}(1-\mu_{it})$. It is shown that $\mathrm{EF}^* = \mathrm{EF}$.

As stated in section 4.3.2,

$$c_{it} = v_{it}(w_{it}^{(1)} - w_{it}^{(0)})$$

$$Var(g_{it}^*) = v_{it}b_{it} \quad and \quad Cov(g_{it}^*, g_{it'}^*) = \rho v_{it}^{1/2} v_{it'}^{1/2} b_{it} b_{it'},$$

where

$$b_{it} = (1-\mu_{it})w_{it}^{(1)} + \mu_{it}w_{it}^{(0)}.$$

Let $\mathrm{B}_i = Diag\{b_{it}\}$. It can be shown that $\mathrm{E}[\partial g_{it}^*/\partial\eta_{it}] = -v_{it}b_{it}$ which implies

$$C_i^* = Diag\{\mathrm{E}[\partial g_{it}^*/\partial\eta_{it}]\}X_i = B_iC_i$$

Thus,

$$\mathrm{EF}^* = \sum_{i=1}^{K} C_i^{*\prime} V_i^{*-1} g_i^* = \sum_{i=1}^{K} C_i' B_i (B_i V_i B_i)^{-1} g_i^* = \sum_{i=1}^{K} C_i' V_i \left(g_{i1}^*/b_{i1}, \ldots, g_{in_i}^*/b_{in_i} \right)'$$

Consider $y_{it} = 0$. Then

$$g_{it}^*(y_{it}=0)/b_{it} = \frac{-w_{it}^{(0)}\mu_{it} - c_{it}}{b_{it}} = -\mu_{it} = g_{it}(y_{it}=0).$$

Similarly,

$$g_{it}^*(y_{it}=1)/b_{it} = \frac{w_{it}^{(1)}(1-\mu_{it}) - c_{it}}{b_{it}} = 1-\mu_{it} = g_{it}(y_{it}=1).$$

Thus, for binary data $g_{it}^*/b_{it} = g_{it}$ and $\mathrm{EF}^* = \mathrm{EF}$.

REFERENCES

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford: Clarendon Press.

Bai, Rao, and Wu (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica* **2**, 237-254.

Barrett, B. E. and Ling, R. F. (1992). General classes of influence measures for multivariate regression, *Journal of the American Statistical Association*, **87**, 184-191.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley.

Bhapkar, V. P. (1972). On a measure of efficiency of an estimating equation. *Sankhya* A **34**, 467-72.

Box, G. E. P. (1953). Non-normality and test on variances. *Biometrika* **40**, 318-335.

Carroll, R. J. and Pederson S. (1993). On robustness in the logistic regression model. *J. R. Statist. Soc. B*, **55**, 693-706.

Carroll, R., and Ruppert (1988). *Transformations and weighting in regression.* Chapman and Hall. New York.

Chatterjee, S., and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, **1**, 379-416.

Christensen, R., Pearson, L. M. and Johnson, W. (1992). Case-deletion diagnostics for mixed models, *Technometrics*, **34**, 38-45.

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata, *Biometrika*, **81**, 413-417.

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

Cook, R. D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association*, **74**, 169-174.

Cook, R. D. (1986). Discussion of paper by Chatterjee and Hadi. *Statistical Science*, **1**, 393-397.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *J. R. Statist. Soc.* B, **50**, 225-265.

Cox, D. R. and Hinkley (1968). A note on the efficiency of least-squares estimates. *J. R. Statist. Soc.* B **30**, 284-9.

Cox, D. R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution, *Biometrika*, **81**, 403-8.

Crowder (1978). Beta-binomial anova for proportions. *Appl. Statist.* **27**, 34-7.

Crowder (1987). On linear and quadratic esitmating functions. *Biometrika* **74**, 591-7.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika* **74**, 233-45.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-12.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* **55**, 231-44.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* New York: Wiley and Sons.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.* A6(9), 813-827.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **36**, 1753-1758.

Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *Ann. of Stat.*, Vol. **1**, No. 5, 799-821.

Huggins, R. M. (1993). A robust approach to the analysis of repeated measures. *Biometrics,* **49**, 715-20.

Johnson, W. (1985). Influence measures for logistic regression: another point of view, *Biometrika*, **72**, 59-65.

Kay, R. and Little, S. (1986). Assessing the fit of the logistic model: a case study of children with the Haemolytic Uraemic Syndrome, *Appl. Statist.*, **35**, 16-30.

Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods, 2nd. ed.*, Boston: PWS-Kent Publishing Company.

Kunsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, **84**, 460-66.

Liang, K.-Y. and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics*, **49**, 623-630.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.

Liang, K.-Y., Zeger, S., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *J. R. Statist. Soc.* B **54**, No. 1, 3-40.

Mancl, L. and Leroux, B. (1995). Efficiency of regression estimates for clustered data. Technical Report, Biometry Core, Regional Clinical Dental Research Center, University of Washington, Seattle.

Marrona, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51-67.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

McCullagh, P. (1991). Quasi-likelihood and estimating functions. In: Hinkley, D., Reid, N. and Snell, E., eds. *Statistical Theory and Modelling*. London: Chapman and Hall.

McCullagh, P and Nelder, J.A. (1989). *Generalized Linear Models, 2nd ed.*, London: Chapman and Hall.

Morgenthaler, S (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, **79**, 747-54.

Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika*, **68**, 227-33.

Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 2, 221-32.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370-84.

Pregibon, D. (1981). Logistic regression diagnostics, *Ann. of Stat.*, **9**, 705-724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics*, **38**, 485-98.

Pregibon, D. (1988). Discussion of paper by Copas. *J. R. Statist. Soc.* B, **50**, 225-265.

Preisser, J., Koch, G., and Shockley, W. (1993). Statistical considerations for cross-sectional data relating to tracheal reconstruction over time. *J. Biopharm. Stat.*, Vol. **3**. No. 2, 167-183.

Prentice, R. and Zhao (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825-39.

Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting. *JASA*, **84**, 373-80.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*, New York: John Wiley.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*, New York: Chapman and Hall.

Serfling, R. J. (1980). *Approximation theorems fo mathematical statistics.* Wiley, New York.

Simpson, D., Carroll, R., and Ruppert, D. (1987). M-estimation for discrete data: asymptotic distribution theory and implications. *Ann. of Stat.*, Vol. **15**, No. 2, 657-669.

Singer, J. M. and Sen, P. K. (1985). M-methods in multivariate linear models. *Journal of Multivariate Analysis*, **17**, 168-84.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-47.

Welsch, R. E. (1986). Discussion of paper by Chatterjee and Hadi. *Statistical Science*, **1**, 403-405.

Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions, *Appl. Statist.*, **36**, 181-191.

Zeger, S. and Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, **11**, 1825-39.

Zeger, S., Liang, K.-Y., and Albert, P. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049-60.

Zeger, S. and Qaqish, B. (1988). Markov regression models for time-series: a quasi-likelihood approach. *Biometrics*, **44**, 1019-31.