1994

The Standardized Influence Matrix and Its Applications to Generalized Linear Models

Jiandong Lu

Virginia Commonwealth University
School of Medicine

This is to certify that the dissertation prepared by Jiandong Lu entitled "The Standardized Influence Matrix and Its Applications in Generalized Linear Models" has been approved by his committee as satisfactory completion of the dissertation requirement for the degree of Doctor of Philosophy.

[redacted]
Daijin Ko, Ph.D., Co-director of Dissertation

[redacted]
Pippa M. Simpson, Ph.D., Co-director of Dissertation
Children's Hospital of Michigan
Wayne State University

[redacted]
Walter H. Carter, Jr., Ph.D., School of Medicine

[redacted]
Linda A. Corey, Ph.D., School of Medicine

[redacted]
Patricia Pepple Williamson, Ph.D., College of Humanities and Sciences

[redacted]
Walter H. Carter, Jr., Ph.D., Department Chairman

[redacted]
Hermes A. Kontos, Ph.D., Dean, School of Medicine

[redacted]
William L. Dewey, Ph.D., Dean, School of Graduate Studies

July 13, 1994
Date

The Standardized Influence Matrix and Its Applications to Generalized Linear Models

Dissertation

submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics at the Medical College of Virginia Virginia Commonwealth University

By

Jiandong Lu

B.S., Fudan University
Shanghai, People's Rep. of China, 1985

M.S., Shanghai Jiaotong University
Shanghai, People's Rep. of China, 1988

Research Director: Dr. Daijin Ko
Dr. Pippa M. Simpson

Department of Biostatistics

Medical College of Virginia
Virginia Commonwealth University
Richmond, Virginia

August, 1994

Acknowledgements

I wish to express my sincere appreciation to my advisors, Dr. Daijin Ko and Dr. Pippa M. Simpson, for their friendship, supervision and encouragement. I would not have done it without them. I would also like to thank the members of my graduate committee, Dr. Walter Hans Carter, Dr. Linda Corey and Dr. Patricia Pepple Williamson, for their valuable time and comments.

A genuine thanks goes to Dr. Sung Choi, Dr. Chris Gennings, Dr. R.K. Elswick and other faculty members. I have greatly benefitted from their experience and knowledge over the past four years.

And I owe special thanks to my wife for her constant love and support throughout my years of study. Finally I dedicate this dissertation to my parents who have sacrificed so much for their children.

Table of Contents

List of Tables

List of Figures

The Standardized Influence Matrix and its Applications to Generalized Linear Models

ABSTRACT

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics at Medical College of Virginia, Virginia Commonwealth University.

Jiandong Lu

Medical College of Virginia

Virginia Commonwealth University

Research Director: Dr. Daijin Ko
Dr. Pippa M. Simpson

The standardized influence matrix is a generalization of the standardized influence function and Cook's approach to local influence. It provides a general and unified approach to judge the suitability of statistical inference based on parametric models. It characterizes the local influence of data deviations from parametric models on various estimators, including generalized linear models. Its use for both robustness measures and diagnostic procedures has been studied. With global robust estimators, diagnostic statistics are proposed and shown to be useful in detecting influential points for linear regression and logistic regression models. Robustness of various estimators is compared via the standardized influence matrix and a new robust estimator for logistic regression models is presented.

Chapter 1

Introduction

The generalized linear model (McCullagh and Nelder 1989) has found wide usage in a variety of fields, especially in biomedical and clinical research, and epidemiology studies. Unfortunately, like all mathematical models, generalized linear models are approximations to the situations to which they are applied. One may try to distinguish three main groups of reasons for it (Hampel 1962):

1. "gross error": something went wrong when the data was recorded;

2. limited accuracy of measurements;

3. the underlying "true" distribution differs from those given in the parametric model.

It is shown that these deviations cannot be ignored in practice, since even very mild and not obvious deviations from the assumed parametric model may change the behavior of the "optimal" estimator drastically for the worse (see Tukey 1960).

There are fundamentally two approaches to tackling the problem of model approximation. One approach is to provide diagnostic tools to judge the appropriateness of the model; while the other is to accommodate a certain degree of discrepancy in the statistical inference (parameter estimation and hypothesis testing) procedures that are robust to some types of model violation. The former is known as a diagnostics approach, and the latter as

a robust statistics approach.

A number of researchers have devoted their efforts to these two approaches for the linear model in the last twenty years. The diagnostics approach attempts to evaluate model assumptions by comparing the fitted model with the observed data. Typically, a measure of influence of a data point on the parameter estimates is comprised of two parts: the residual and the leverage. A large value of either the residual or the leverage indicates that the corresponding data point has undue influence on the statistical procedure. On the other hand, robust statistical procedures adaptively assign weight to each data point and then obtain the parameter estimates, where a small weight implies that the corresponding data point is likely to be contaminated.

Apparently many issues still remain unanswered. Two of them are: joint influence assessment in regression and discovering problems from specific models.

1.1 Joint Influence in Regression

The developments discussed above deal with the influence of a single data point. In particular, the concept of *Gross-error sensitivity*, defined as the supremum of the Euclidean norm of the influence function, measures the worst influence of an estimator. However, the influence of a group of the data points may not be simply the summation of the influences of all the single data points in the group. For example, when two contamination points are on one side of the regression line, the joint influence of these two points is greater than either single influence and is approximately the sum of both single influences if the two points are sufficiently far away from the line. Moreover, when two contamination points are on the opposite sides of the regression line, the joint influence of these two points is less than either single influence and may be canceled out if the two points are symmetric with

respect to the line.

Therefore, a measure of joint influence of a group of data point is needed so that it is not only compatible with single influence but also is effective in finding the group of data points that have joint undue influence on the parameter estimates.

1.2 Generalized Linear Model

Many of the approaches applied to linear models can be extended to generalized linear models. However, the differences of the model settings trigger some unique problems for specific generalized linear models including logistic regression and survival time regression models. For instance, the distribution of residuals in logistic regression is no longer free from parameters; some robust technique that is useful in linear regression may not work because of the discrete distribution; censoring observations make it hard to define the residuals, and consequently the concept of outlier needs to be revised in survival time regression models. Therefore, more research dealing with specific problems is essential to the success and effectiveness of diagnostic and robust approaches.

1.3 Prospectus

This dissertation introduces the standardized influence matrix and uses it as a basic tool to assess the joint local influence of various kinds of estimators in regression. The standardized influence matrix is essentially a quadratic form of the influence function of parameter estimates evaluated at a group of observations in a regression model. By studying the standardized influence matrix of an estimator (e.g. least squares estimator, M-estimator) in a regression model, we can examine the robustness of the estimator in terms of joint local influence. In addition, we can assess the joint local influence of a group of cases

on the inferences of the model and identify the influential observations. Equipped with the tool we propose, we give extensive discussion of the estimators of linear and logistic regression model. Moreover, for purposes of diagnostic applications, some of the global robust estimators are studied.

Chapter 2 presents the review and a brief discussion of existing techniques for influence assessment in generalized linear models. Chapter 3 proposes a standardized influence matrix and gives its properties. Contributions to robust statistics and diagnostics approaches are given in chapter 4. Chapter 5 and chapter 6 describe applications in linear regression and logistic regression respectively. Chapter 7 discusses a few unique issues of logistic regression for binary responses, and presents a new global robust estimator for binary logistic regression. Results of simulation studies using this estimator are given in chapter 8. In chapter 9, the proposed technique is used in the analysis of a few datasets for both linear regression and logistic regression. In chapter 10 conclusions are drawn and future work is discussed.

Chapter 2

Influence Assessment in Regression

A practical and well-established approach to influence assessment in regression is based on case deletion. Recently the idea of small perturbations and methods based on robust regression estimators have become the two main vehicles for assessing influence of multiple observations. This chapter reviews deletion, Cook's local influence and robust regression approaches to the problem of multiple influential observations in linear regression. Section 2.4 introduces diagnostic methods for the generalized linear model. And finally Section 2.5 discusses the advantage and weakness of each approach and then suggests that certain measures of influence incorporated with robust estimators be a solution to the problem.

2.1 Deletion Approach

The deletion approach examines how the various measures involved in a regression analysis of the data change when some of the observations are deleted. A number of measures based on various quantities have been discussed by Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Chatterjee and Hadi (1988). This section introduces some of those measures that are well known and their extensions to the case of multiple influential observations.

The model for linear regression is defined as

$$Y = X\beta + \varepsilon \tag{2.1.1}$$

and

$$E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2 I$$

where X is an $n \times p$ full rank matrix of known covariates, Y is an n-vector of observed responses, β is a p-vector of unknown parameters, and ε is an n-vector of unobservable error with indicated distributional properties.

The main results based on the least square estimator are summarized by

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \tag{2.1.2}$$

$$H = X(X^T X)^{-1} X^T, \tag{2.1.3}$$

$$e = Y - \hat{Y} = (I - H)Y, \tag{2.1.4}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p} e^T e = \frac{1}{n-p} Y^T (I - H) Y. \tag{2.1.5}$$

The notations of $Y_{(I)}$, $X_{(I)}$, $\sigma_{(I)}$, where $I = \{i_1, i_2, \ldots, i_m\}$ is a subset of $\{1, 2, \ldots, n\}$, represent the corresponding quantities when the m observations indexed by I are deleted.

2.1.1 Single Deletion

1. Residuals and Hat Matrix

   Residuals play an important role in regression diagnostics; no analysis is complete without a thorough examination of the residuals.

   (a) The ordinary residuals and hat matrix are

$$e = Y - \hat{Y} = (I - H)Y$$

$$= (I - H)\varepsilon \tag{2.1.6}$$

and

$$Var(e) = \sigma^2(I - H). \tag{2.1.7}$$

Those equations demonstrate clearly that the relationship between e and ε depends on the hat matrix H. It is well known that the diagonal element in the hat matrix, h_{ii}, called the *leverage*, is a useful indicator of whether or not the ith case is outlying with respect to its X values. If the h_{ij}'s are sufficiently small, e will serve as a reasonable substitute for ε, otherwise the usefulness of e may be limited.

(b) Studentized Residual

The ordinary residuals have a distribution that is scale dependent since the variance of each e_i is a function of both σ^2 and h_{ii}. It is useful to define a studentized version of the residuals that does not depend on either of these quantities.

- Internally studentized residual

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \ i = 1, 2, \ldots, n. \tag{2.1.8}$$

Ellenberg(1973) showed that if $rank(X_{(i)}) = p$, then each

$$\frac{r_i^2}{n - p}, \ i = 1, 2, \ldots, n$$

is identically distributed as Beta$(\frac{1}{2}, \frac{1}{2}(n - p - 1))$.

- Externally studentized residual

$$r_i^* = \frac{e_i}{\hat{\sigma_{(i)}}\sqrt{1-h_{ii}}}, \; i = 1, 2, \ldots, n. \tag{2.1.9}$$

Under normality, $\hat{\sigma}^2_{(i)}$ and e_i are independent, and r_i^* follows a Student's t distribution with $n - p - 1$ degrees of freedom.

The relationship between r_i and r_i^* is

$$r_i^* = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}, \tag{2.1.10}$$

which shows that r_i^* is a monotonic transformation of r_i.

2. Mean Shift Outlier Model

A worthwhile framework used to study outliers is the mean shift outlier model,

$$Y = X\beta + d_i\lambda + \varepsilon,$$

and

$$E(\varepsilon) = 0, \; Var(\varepsilon) = \sigma^2 I \tag{2.1.11}$$

where d_i is an n-vector with i-th element equal to 1, and all other elements equal to zero. Nonzero values of λ imply the i-th case is an outlier.

It can be shown that the F-test statistic for $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ is identical to the externally studentized residual. However, there are two arguments made by Cook and Weisberg (1982), Chatterjee and Hadi (1988) presenting the limitations.

- F_i, which is equal to r_i^{*2}, follows a noncentral F-distribution with noncentrality parameter $\lambda^2(1 - h_{ii})/\sigma^2$. For large h_{ii}, the noncentrality parameter is small; hence it will be difficult to distinguish between the distributions of F_i under

H_0 and under H_1. Finding outliers at remote points will be more difficult than finding outliers where h_{ii} is small. Also, since h_{ii} is increasing in p, outliers become more difficult to detect as the number of model parameters increases.

- When the candidate case for an outlier is unknown, the test is usually based on the maximum of the r_i^{*2} over all i. A multiple testing procedure must be used to find significance levels. Several procedures and rules are proposed and discussed by Cook and Prescott (1981), Lund (1975), Prescott (1975) and Tietjen et al (1973).

3. Influence on the regression coefficients

An overall measure of the combined impact of the ith case on all of the estimated regression coefficients is *Cook's distance* D_i. This measure is derived from the concept of a confidence region for all p regression coefficients simultaneously. It can be shown that the boundary of this confidence region for multiple regression model (2.1.1) is given by:

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{p\hat{\sigma}^2} = F(1-\alpha, p, n-p). \tag{2.1.12}$$

Cook's distance D_i uses the same structure for measuring the combined impact of the *differences in the estimated regression coefficients* when the ith case is deleted:

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})^T X^T X(\hat{\beta}-\hat{\beta}_{(i)})}{p\hat{\sigma}^2} \tag{2.1.13}$$

$$= \frac{e_i^2}{p\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]. \tag{2.1.14}$$

Note that D_i depends on two factors: the size of residual e_i and the leverage value h_{ii}. Thus either large e_i or large h_{ii} will result in large D_i, and the ith case may be influential because of a large value of residual or of leverage.

It can be shown that Cook's Distance is a measure based on a sample version of the influence function of the regression parameters. Also Chatterjee and Hadi (1988) has summarized some other influence measures based on different kinds of influence functions.

4. Likelihood distance

Let $l(\beta, \sigma^2)$ be the log-likelihood function based on all n observations. Let $\tilde{\beta}$ and $\tilde{\sigma}^2$ be the maximum likelihood estimate of β and σ^2, respectively, and $l(\tilde{\beta}, \tilde{\sigma^2})$ be the log-likelihood evaluated at $\tilde{\beta}$ and $\tilde{\sigma}^2$. It is well known that the $100(1-\alpha)\%$ asymptotic confidence region for β and σ^2 is given by

$$\{(\beta, \sigma^2) : 2[l(\tilde{\beta}, \tilde{\sigma}^2) - l(\beta, \sigma^2)] \leq \chi^2_{\alpha;p+1}\}.$$

Under the normality assumption, it follows that

$$\tilde{\beta} = \hat{\beta} \tag{2.1.15}$$

and

$$\tilde{\sigma}^2 = \hat{\sigma}^2 \left(\frac{n-p}{n}.\right) \tag{2.1.16}$$

When the ith observation is deleted, the MLE of β and σ^2 are

$$\tilde{\beta}_{(i)} = \hat{\beta}_{(i)} \tag{2.1.17}$$

and

$$\tilde{\sigma}^2_{(i)} = \hat{\sigma}^2_{(i)} \left(\frac{n-p-1}{n-1}\right). \tag{2.1.18}$$

The influence of the ith observation on the likelihood function may be measured by the distance between the likelihood functions evaluated at $(\tilde{\beta}, \tilde{\sigma}^2)$ and $(\tilde{\beta}_{(i)}, \tilde{\sigma}^2_{(i)})$. Cook

and Weisberg (1982) define the likelihood distance by

$$LD_i(\beta, \sigma^2) = 2[l(\tilde{\beta}, \tilde{\sigma}^2) - l(\tilde{\beta}_{(i)}, \tilde{\sigma}^2_{(i)})],\ i = 1, 2, \ldots, n. \tag{2.1.19}$$

It can be shown that under the normality assumption, $LD_i(\beta, \sigma^2)$ becomes

$$\begin{aligned} LD_i(\beta, \sigma^2) &= n \ln \frac{n(n-p-r_i^2)}{(n-1)(n-p)} \\ &\quad + \frac{(n-1)r_i^2}{(1-h_{ii})(n-p-r_i^2)} - 1, \end{aligned} \tag{2.1.20}$$

where r_i is the internally studentized residual as defined in (2.1.8). Note that $LD_i(\beta, \sigma^2)$ is useful if we are interested in estimating both β and σ^2. If we are only interested in estimating β, the likelihood distance is defined as

$$LD_i(\beta|\sigma^2) = 2[l(\tilde{\beta}, \tilde{\sigma}^2) - \max_{\sigma} l(\tilde{\beta}_{(i)}, \sigma^2)],\ i = 1, 2, \ldots, n. \tag{2.1.21}$$

And it becomes

$$LD_i(\beta|\sigma^2) = n \ln \left(1 + \frac{p}{n-p} D_i\right),\ i = 1, 2, \ldots, n \tag{2.1.22}$$

where D_i is given by (2.1.14). Thus $LD_i(\beta|\sigma^2)$ is equivalent to Cook's distance.

5. Graphical displays

Atkinson (1981) presented half normal plots with envelopes derived from simulation to detect the outlying and influential observations in regression. Atkinson's display is basically a quantile plot or a half normal plot for certain influence measures, such as studentized residual or Cook's distance, with a simulated envelope roughly like a simultaneous confidence region of the measure.

Take the squared root of Cook's distance as an example, for a problem with a fixed hat matrix H, m pseudo-random n-vectors $\varepsilon_1, \ldots, \varepsilon_m$ generated from a standard normal distribution. The pseudo squared root of Cook's Distance d_k, $k = 1, \ldots, m$ are then

computed. Let the ordered elements of d^k be denoted by $d_{(i)k}$ and, for each i, let $d^{\eta}_{(i)}$, $0 < \eta < 1$, denote the $100 \times \eta$ percentile of the empirical distribution of $\{d_{(i)k},\ k = 1, \ldots, m\}$. Simultaneous plots of two n-vectors with elements $(d^{\alpha/n}_{(i)})$ and $(d^{1-\alpha/n}_{(i)})$ is the so-called envelope. If the observed squared root of Cook's distances fall beyond or near the boundary of the envelope, the corresponding observations are considered influential.

2.1.2 Multiple Deletion

Almost all single deletion approaches except the graphical display method can be generalized to multiple observations measures.

Let $I = \{i_1, i_2, \ldots, i_m\}, m < n - p$, be the set containing the indices of the m observations to be deleted. Without loss of generality, let us assume that the m observations are the last m observations, so that Y, X can be written as

$$Y = \begin{pmatrix} Y_{(I)} \\ Y_I \end{pmatrix} \begin{matrix} (n-m) \times 1 \\ m \times 1 \end{matrix},$$

$$X = \begin{pmatrix} X_{(I)} \\ X_I \end{pmatrix} \begin{matrix} (n-m) \times p \\ m \times p \end{matrix}.$$

The following is a summary of the generalization of single deletion approaches:

- The generalization of the square of the internally studentized residual becomes

$$r_I^2 = \frac{e_I^T (I - H_I)^{-1} e_I}{\hat{\sigma}^2} \tag{2.1.23}$$

where

$$e_I = Y_I - X_I^T \hat{\beta} \tag{2.1.24}$$

$$H_I = X_I^T (X^T X)^{-1} X_I. \tag{2.1.25}$$

Similarly, the externally studentized residual is formed from the internally studentized residual with $\hat{\sigma}^2$ replaced by $\hat{\sigma}^2_{(I)}$.

- The generalized mean shift outlier model is given by

$$Y = X\beta + U_I\phi + \varepsilon$$

$$E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2 I, \tag{2.1.26}$$

where

$$U_I = \begin{pmatrix} 0 \\ I \end{pmatrix} \begin{matrix} (n-m)\times m \\ m \times m \end{matrix} \tag{2.1.27}$$

and ϕ is an $m \times 1$ vector. A nonzero vector of ϕ implies that the m observations indexed by I are outliers. Similarly the F-test statistic

$$F_I = \frac{e_I^T(I - H_I)^{-1}e_I}{m\hat{\sigma}^2_{(I)}} \tag{2.1.28}$$

is identical to the corresponding externally studentized residual.

- Generalized Cook's distance becomes

$$D_I = \frac{e_I^T(I - H_I)^{-1}H_I(I - H_I)^{-1}e_I}{p\hat{\sigma}^2}. \tag{2.1.29}$$

- Generalized likelihood distance becomes

$$\begin{aligned} & LD_I(\beta, \sigma^2) \\ &= 2[l(\tilde{\beta}, \tilde{\sigma}^2) - l(\tilde{\beta}_{(I)}, \tilde{\sigma}^2_{(I)})] \\ &= n\ln\left(\frac{n(n-p-r_I^2)}{(n-m)(n-p)}\right) + \frac{(n-m)(n-p+pD_I)}{n-p-r_I^2} - n. \end{aligned} \tag{2.1.30}$$

2.2 Differentiation Approach and Local Influence

We now examine a second means for identifying influential observations: differentiation of regression outputs with respect to specific perturbations.

2.2.1 Ordinary Approach

Instead of deleting the ith observation, we can alter the weight attached to the ith observation. In model (2.1.1), we replace $var(\epsilon_i) = \sigma^2$ with $var(\epsilon_i) = \sigma^2/\omega_i$, for the specific i only. Differentiation of the regression coefficients with respect to ω_i, evaluated at $\omega_i = 1$, provides a means of examining the sensitivity of the regression coefficients to a slight change in the weight given to the ith observation. Large values of this derivative indicate observations that have large influence on the calculated coefficients. It can be shown that

$$\frac{\partial \hat{\beta}(\omega_i)}{\partial \omega_i} = \frac{(X^T X)^{-1} \boldsymbol{x}_i^T e_i}{[1-(1-\omega_i)h_{ii}]^2}, \tag{2.2.1}$$

and it follows that

$$\Delta\hat{\beta}_i = \left.\frac{\partial \hat{\beta}(\omega_i)}{\partial \omega_i}\right|_{\omega_i=1} = (X^T X)^{-1} \boldsymbol{x}_i^T e_i. \tag{2.2.2}$$

The last formula is often viewed as the influence of the ith observation on the estimated coefficients. And we can construct statistics similar to Cook's distance as

$$F_i = \frac{\Delta\hat{\beta}_i^T (X^T X) \Delta\hat{\beta}_i}{p\hat{\sigma}^2} \tag{2.2.3}$$

$$= \frac{h_{ii} e_i^2}{p\hat{\sigma}^2}. \tag{2.2.4}$$

Obviously the idea can be easily generalized to multiple weights cases. However, the interpretation of the matrix instead of the scaler is more difficult. The eigenvalues and eigenvectors approach is thoroughly studied in Cook's local influence approach.

2.2.2 Local Influence

In his paper, Cook (1986) attempts to "present a general method for assessing the local influence of minor perturbations of a statistical model". The approach relies on the concept of likelihood displacement and certain elementary ideas from differential geometry.

For a given set of observed data, let $L(\theta)$ denote the log-likelihood corresponding to the postulated model, where θ is a $p \times 1$ vector of unknown parameters. Perturbations are introduced into the model through the $q \times 1$ vector ω which is restricted to some open subset Ω of R^q. ω can reflect any well-defined perturbation scheme. For instance, ω can be a collection of case weights, which is the situation of interest in this dissertation. Now let $L(\theta \mid \omega)$ denote the log-likelihood corresponding to the perturbed model for a given ω in Ω. We assume that there exists an ω_0 such that $L(\theta) = L(\theta \mid \omega_0)$ for all θ. Finally, let $\hat{\theta}$ and $\hat{\theta}_\omega$ denote the maximum likelihood estimators under $L(\theta)$ and $L(\theta \mid \omega)$ respectively.

To assess the influence of ω throughout Ω, Cook and other researchers consider the likelihood displacement

$$LD(\omega) = 2[L(\hat{\theta}) - L(\hat{\theta_\omega})]. \tag{2.2.5}$$

Then he introduces the influence graph $\alpha(\omega)$ which is formed by the value of the $(q+1) \times 1$ vector

$$\alpha(\omega) = \begin{pmatrix} \omega \\ LD(\omega) \end{pmatrix} \tag{2.2.6}$$

as ω varies throughout Ω. Ideally we would like to view a complete influence graph to assess influence in a particular problem. Clearly, this is possible only in simplest situations so that it is necessary to consider methods for extracting the information contained in an influence graph. Thus Cook suggests a local measure, geometric normal curvature, of influence for characterizing the behavior of an influence graph around $\omega = \omega_0$.

The basic form for normal curvature is

$$C_l = 2|l^T \ddot{F} l|, \tag{2.2.7}$$

where l is a nonzero vector of unit length and $\ddot{F}$ is a $q \times q$ matrix with elements $\partial^2 L(\hat{\theta}_\omega)/\partial\omega_k \partial\omega_j$, $j, k = 1, ..., q$. There are several ways of using the normal curvature to

study the influence in practice. The extremes $C_{max} = max_l C_l$ and $C_{min} = min_l C_l$, which correspond to the maximum and minimum absolute eigenvalues of $\ddot{F}$, are two possible options. The eigenvector l_{max} associated with C_{max} indicates how to perturb the postulated model to obtain the greatest local change in the likelihood displacement. In the situation of case weights, the relatively large ith element of l_{max} indicates that perturbations in the weight ω_i of the ith case may lead to substantial changes in the result of the analysis and thus that ω_i is relatively influential.

Cook then applied the approach in normal linear regression models. The following are the main results of his application:

- Individual cases.

  The curvature for the influence graph obtained by modifying the weight attached to a single case, say the ith, is

  $$C = 2e_i^2 h_{ii}/\hat{\sigma}^2 = 2p(1 - h_{ii})^2 D_i \tag{2.2.8}$$

  where D_i is Cook's distance as defined in (2.1.13-14). This equation is identical to the ordinary differentiation approach. Then Cook remarks that "C_l measures the influence of local changes in the case weight, while D_i measures global changes ".

- Multiple cases

  Let ω be the n-vector of case weights for the regression model (2.1.1) and assume that σ^2 is known. After differentiating the log-likelihood $L(\beta|\omega)$ with respect to β and ω, and evaluating at $\hat{\beta}$ and $\omega_0 = 1$, he has

  $$C_l = 2l^T D(e) \cdot H \cdot D(e) l/\sigma^2, \tag{2.2.9}$$

  where e and H are the ordinary residuals and hat matrix respectively, and $D(e) = \text{diag}(e_1, \ldots, e_n)$. Furthermore, when σ^2 is unknown and if only β is of interest, it can

be shown that the curvature is given by the above equation with σ^2 replaced with $\hat{\sigma}^2$. He also gives a few examples to illustrate the behaviors of C_{max} and l_{max}.

- General reference for C_{max}.

  Cook uses a simple random sample as a rationale to decide a curvature of 2 serves as a useful general reference. Curvature much larger than 2 indicates notable local sensitivity. Since it is quite controversial to treat 2 as a criterion for being influential, he stressed that "regardless of the size of C_{max}, an inspection of l_{max} is worthwhile."

2.3 Robust Regression Methods

The third approach to influence assessment is to present measures based on robust parameter estimates. Both Atkinson (1986) and Rousseeuw and Zomeren (1990) use the least median square estimator as a basic tool to construct the influence measures for each individual observation.

- In Atkinson's paper the least median square estimator serves as an exploratory tool for the identification of outliers. The techniques of multiple deletion regression diagnostics are then directly applicable for confirmation of the presence of outliers and influential observations.

- Since an influential observation can be either an outlier or a high leverage point, Rousseeuw and Zomeren suggest using robust distances as a measure of leverage and standardized least median residuals as an indication of outlying. The feature of their paper is to use estimators with a high breakdown point (see chapter 4 for the definition), particularly minimum volume ellipsoid estimator (MVE) and least median square estimator (LMS), to obtain the distances for multivariate independent variables

and the residuals of the regression model. The graphic display in which the robust residuals versus the robust distances, classifies the data into regular observations, vertical outliers, good leverage points, and bad leverage points.

2.4 Generalized Linear Model and Logistic Regression

In this section, we study the proposed diagnostic methods for a generalized linear model (GLM), concentrating on logistic regression in particular. In a GLM we assume the following:

1. $y_i, i = 1, 2, \ldots, n$, are independently distributed with expected value μ_i and come from a probability density belonging to the exponential family $f(y_i, \theta_i, \phi)$ where

$$f(y_i, \theta_i, \phi) = \exp\{[\theta_i y_i - b(\theta_i)]/a(\phi) + c(\alpha_i, y_i)\} \tag{2.4.1}$$

and the function $a(\phi)$ is commonly of the form

$$a(\phi) = \phi/m.$$

ϕ is called the *dispersion parameter*. The equation (2.4.1) characterizes the random component of the model,

2. The covariates $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip})$ provide a linear predictor,

$$\eta_i = \boldsymbol{x}_i\beta, \tag{2.4.2}$$

of μ_i and constitute the systematic component of the model.

3. The link between the systematic and random components is given by

$$\eta_i = g(\mu_i) \tag{2.4.3}$$

where g is any monotonic differentiable function of μ. A link function in which $\eta_i = \mu_i$ is called the canonical link function.

Except for the case of normal errors with identity link function, an iterative scheme is needed to obtain the maximum likelihood estimators of β. As Nelder and Wedderburn (1972) pointed out, the solution of maximum likelihood equations with canonical link function is equivalent to an iterative weighted least squares procedure with a weight function

$$w_i = v_i^{-1}\left(\frac{d\mu_i}{d\eta_i}\right)^2, \tag{2.4.4}$$

where

$$v_i = Var(y_i) = b''(\theta_i)a(\phi). \tag{2.4.5}$$

This leads to the algorithm

$$\hat{\beta}^{t+1} = \hat{\beta}^t + (X^T\hat{W}X)^{-1}X^TS, \tag{2.4.6}$$

where $\hat{W} = \text{diag}(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n)$, $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)^T$ and $S = Y - \hat{\mu}$. We define the analogue hat matrix

$$\hat{H} = \hat{W}^{1/2}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{1/2}. \tag{2.4.7}$$

In the situation of binary logistic regression, y_i follows a binomial distribution with parameters 1 and p_i,

$$\begin{aligned} f(y_i, p_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \exp\{y_i \ln(\frac{p_i}{1-p_i}) + \ln(1-p_i)\}. \end{aligned} \tag{2.4.8}$$

Thus,

$$\theta_i = \ln(\tfrac{p_i}{1-p_i}), \tag{2.4.9}$$

$$b(\theta_i) = -\ln(1-p_i) = \ln(1+\exp(\theta_i)), \tag{2.4.10}$$

$$\mu_i = p_i = \tfrac{1}{1+\exp(-\theta_i)}, \tag{2.4.11}$$

$$v_i = Var(y_i) = p_i(1-p_i), \tag{2.4.12}$$

$$w_i = p_i(1-p_i), \tag{2.4.13}$$

and the link function for logistic regression is

$$\eta = g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right). \tag{2.4.14}$$

2.4.1 Residuals

For a generalized linear model, we require an extended definition of residuals, applicable to all distributions. It is convenient if these residuals can be used for the same purposes as the normal residuals in section 2.1.

- Pearson residuals

  The Pearson residual, defined by

  $$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{v_i}}, \tag{2.4.15}$$

  is just the raw residual scaled by the estimated standard deviation of Y. And the sum of the squares of the residuals leads to a measure of the discrepancy of the model, which is a generalized Pearson χ^2 statistic. Thus for binary logistic regression, we have

  $$r_{Pi} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}. \tag{2.4.16}$$

- Deviance residuals

  Deviance is another important measure of discrepancy of the model, which takes the form

  $$D = 2\phi[l(Y;Y) - l(\hat{\mu};Y)]. \tag{2.4.17}$$

  Each unit contributes a quantity d_i to that measure, so that $\sum d_i = D$. Hence we define

  $$r_{Di} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}. \tag{2.4.18}$$

For a binary logistic model, we have

$$r_{Di} = \text{sign}(y_i - \hat{p}_i)[2(y_i \ln(y_i/\hat{p}_i) + (1 - y_i)\ln((1 - y_i)/(1 - \hat{p}_i))]^{1/2}. \quad (2.4.19)$$

2.4.2 Single Deletion and One Step Approximation

Since the maximum likelihood estimates (MLEs) for most GLMs are obtained by iterative methods, the MLEs from $n - 1$ cases cannot be obtained as explicit functions of the results of a fit to all n cases. In an important paper Pregibon (1981) derives useful one step approximations for the changes in the MLE and the deviance of the model when the model is perturbed by case weights, and he discusses some diagnostic methods which use these approximations. Williams (1987) describes some GLM diagnostics all of which make use of Pregibon's one step approximations for the change in the components of the deviance when a single case is deleted. The following results are from Williams's paper:

1. All of the diagnostics proposed in his paper are functions of $h_{ii}, r^*_{Pi}, r^*_{Di}$, where h_{ii} is the ith diagonal element of H, r^*_{Pi} is the standardized Pearson residual, which is given by

$$r^*_{Pi} = \frac{r_{Pi}}{\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{v_i(1 - h_{ii})}}, \quad (2.4.20)$$

and r^*_{Di} is the standardized deviance residual, which is given by

$$r^*_{Di} = \frac{r_{Di}}{\sqrt{\phi(1 - h_{ii})}}. \quad (2.4.21)$$

2. One step approximations following single case deletion are considered:

After taking a one step of the iterative weighted least squares procedure with $n - 1$ cases, he obtains

$$\hat{\beta}_{(i)} \approx \hat{\beta} - w_i^{1/2}(1 - h_{ii})^{-1/2} r^*_{Pi}(X^T W X)^{-1} x_i. \quad (2.4.22)$$

Using this approximation, he derives the following:

(a) The decrease in $\sum_{j\neq i} d_j^2$ is approximately $\phi h_{ii} r_{Pi}^{*2}$.

(b) The increase in d_i^2 is approximately $\phi h_{ii}(2-h_{ii})(1-h_{ii})^{-1} r_{Pi}^{*2}$.

(c) The increase in $D = \sum_j d_j^2$ is approximately $\phi h_{ii}(1-h_{ii})^{-1} r_{Pi}^{*2}$.

3. A mean shift outlier model is developed.

The possibility that the ith case is an outlier can be assessed by testing the hypothesis that $\lambda_i = 0$ in the model $\eta = X\beta + U_i\lambda_i$. The likelihood ratio statistic is the reduction G_i in the scaled deviance $\phi^{-1}D$ when the ith case is deleted. Calculation of G_i for each i is computationally expensive. However G_i can be approximated by the contribution of deviance from case i, which is $\phi^{-1}d_i^2$, and that from the remaining $n-1$ cases, which is approximated by $h_{ii}r_{Pi}^2$. Thus

$$G_i \approx r_{Gi}^2 = \phi^{-1}d_i^2 + h_{ii}r_{Pi}^{*2} = (1-h_{ii})r_{Di}^{*2} + h_{ii}r_{Pi}^{*2}. \tag{2.4.23}$$

The $\max r_{Gi}^2$ provides a statistic for testing for the presence of a single outlier. No size of the test is given unless the distribution of y_i is close to Normal. A generalized Atkinson's simulated envelope (see 2.1.1) is useful here to detect the single outlier.

4. Cook's distance is approximated.

Using the one step approximation, Cook's distance is approximated by

$$D_i = \frac{h_{ii}r_{Pi}^{*2}}{p(1-h_{ii})}, \tag{2.4.24}$$

where p is the number of parameters.

Although all the above can be easily extended to the multiple deletion case, the quality of the one step approximation when multiple observations are deleted is not as apparent as that for single deletion. Much more work is required for subset diagnostics.

2.4.3 Pregibon's Differentiation Approach and Cook's Local Influence

Besides the contribution of the one step approximation, Pregibon (1981) introduced an infinitesimal approach to case diagnostics. In logistic regression particularly, the perturbed log-likelihood is expressed as

$$l_\omega(X\beta; Y) = \sum_{j=1}^{n} \omega_j l(\boldsymbol{x}_j\beta; y_j) \tag{2.4.25}$$

where

$$\omega_i = \begin{cases} \omega & \text{for } j = i \\ 1 & \text{otherwise} \end{cases}$$

with $0 < \omega < 1$. The Newton-Raphson method (2.4.6) leads to this one step estimate,

$$\hat{\beta}^1(\omega) = \hat{\beta} - \frac{(X^TWX)^{-1}\boldsymbol{x}_i s_i(1-\omega)}{(1-(1-\omega)h_{ii})}. \tag{2.4.26}$$

Then

$$\frac{\partial}{\partial\omega}\hat{\beta}^1(\omega) = \frac{(X^TWX)^{-1}\boldsymbol{x}_i s_i}{(1-(1-\omega)h_{ii})^2}. \tag{2.4.27}$$

This equation provides the ratio of the changes concerning the effect of the ith observation on the fit. As we mentioned earlier for the deletion approach, the extension to the multiple observation case requires more work.

Since Cook's approach to local influence is based on likelihood displacement, it is applicable to generalized linear models. A few works have been published since then. Thomas and Cook (1989, 1990) applied local influence methods to generalized linear models, while Pettitt and Bin Daud (1989) and Weissfeld (1990) did the same for Cox proportional hazard models. Beckman, Nachtsheim, and Cook (1987) developed and described applications to mixed model analysis of variance. Escobar and Meeker (1988, 1992) described methods for local influence analysis with censored data and parametric regression models.

2.4.4 Copas's Transposition Probability Model

Some special remarks on influence assessment in binary regression models were discussed by Copas (1988). In binary regression, unlike in ordinary linear regression, contamination in y can only take the very simple form of a transposition between 0 and 1. Suppose that such transpositions happen with a small probability γ, so that the actual recorded response y is governed by a probability p^* instead of p, where

$$\begin{aligned} p^* &= (1-\gamma)p + \gamma(1-p) \\ &= p + \gamma(1-2p) \end{aligned} \tag{2.4.28}$$

and p is given by (2.4.11) as before. Independence between the regression and random misclassification parts of the model is assumed, as is independence between different cases. Under the model, an uplier, i.e. $y = 1$ and p near zero, can be explained as a transposition error with probability $p^* = \gamma > 0$ rather than an extremely unusual response from the basic model.

Now the log-likelihood from (2.4.28) is

$$l(\theta, \gamma) = \sum_{i=1}^{n} \{y_i \log p_i^* + (1-y_i)\log(1-p_i^*)\}. \tag{2.4.29}$$

Taking logistic regression, assume γ is sufficiently small for all terms in γ^2 to be ignored, the corresponding score function turns out to be

$$s(\theta, \gamma) = (1+4\gamma)\sum_{i=1}^{n} r_i^* x_i - \gamma \sum_{i=1}^{n} \frac{r_i^* x_i}{p_i(1-p_i)} \tag{2.4.30}$$

where $r_i^* = y_i - p_i^* = y_i - p_i + \gamma(2p_i - 1)$. Then for a specific value of γ, the maximum likelihood estimate $\hat{\theta}_\gamma$ is obtained by setting $s(\theta, \gamma)$ to zero.

Based on $\hat{\theta}_\gamma$, Copas discussed the changes in $\hat{\theta}_\gamma$ as γ increases from zero, likelihood inference for values of γ and diagnostics for individual observations. For details and examples, see Copas (1988) section 4 through 6.

2.5 Influence Assessment and Robust Estimators

Influence assessment is the study of variation in the results of an analysis when the problem formulation is modified. In the regression setting, the perturbation can be perturbations to assumptions, perturbations to data values and perturbations to case weights (including deletion with its zero weight). However, these three sorts of perturbations are related to each other.

From an infinitesimal point of view, perturbations to data values and to case weights are two examples of small deviations from the underlying distribution. The influence function is mainly a heuristic tool to investigate the infinitesimal behavior of a functional on a probability model. Some norms of the influence function will be measures of influence of the contamination.

Single deletion approaches are straight forward and well established. They work well if at most one observation is outlying. Although sequential deletion procedures can detect some of the multiple influential observations, they fail in the presence of masking effects. Hence, multiple deletion approaches are needed to detect joint influence. However, because in most of the cases we have no idea of how many or which observations are candidates as influential points, multiple deletion approaches face massive computational expenses and eventually are hardly useful.

The important feature of Cook's approach to local influence is that it can simultaneously handle perturbations to all cases for linear models, as well as generalized linear models. We can find the cases of most joint influence by looking at maximum curvature and its associated l_{max}. However, the quality of the assessment may be strongly influenced by a few outlying observations. As we have seen, the local influence approach is based on likelihood displacement. One of the problems is that the likelihood, based on all n observations or the

empirical distribution, represents the postulated model only when no outliers exist. Hence with a few outlying observations, the likelihood displacement can be an influence measure of the contaminated model instead of the true model which we are studying. In some sense, local influence approach is just a version of Cook's distance, which fails to detect influential observations in the presence of masking. The following example demonstrates its failure to detect two highly influential points (#16, #17) with masking effect. The data set, shown in Fig.2.1(a), contains 17 observations, where the first 15 observations are generated by

$$y_i = 10 \times x_i + \epsilon_i$$

where $x_i \sim N(0,1)$, and $\epsilon_i \sim N(0, 0.5)$ for $i = 1, ..., 15$. But #16, #17 are (100,-10), (102,-10.2) respectively. Clearly #16 and #17 are high leverage points which are highly influential. Fig.2.1(b), a plot of Cook's Distance versus case number, shows that #9 and #3 are the most influential

points and #16 and #17 have very little influence. Similarly, Fig.2.1(c), a plot of l_{max} versus case number by Cook's approach to local influence, shows that #9 and #3 could be jointly influential with $C_{max} = 77.3$, whereas #16 and #17 have almost the least influence.

Obviously, we can detect #16 and #17 by simply looking at Fig.2.1(a). However, the detection of masking influential points in case of multiple regression would be much harder because it is impossible for us to view a multi-dimensional cloud of data. Therefore, the importance of the example is not for those 17 points per se, but to demonstrate the failure of approaches to local influence based on maximum likelihood estimates.

To unmask the masking influential points, robust estimates, which reflect the majority of the data, should be used to assess influence. In their paper, Rousseeuw and Zomeren (1990) suggest using distances based on high breakdown estimators to detect outliers in

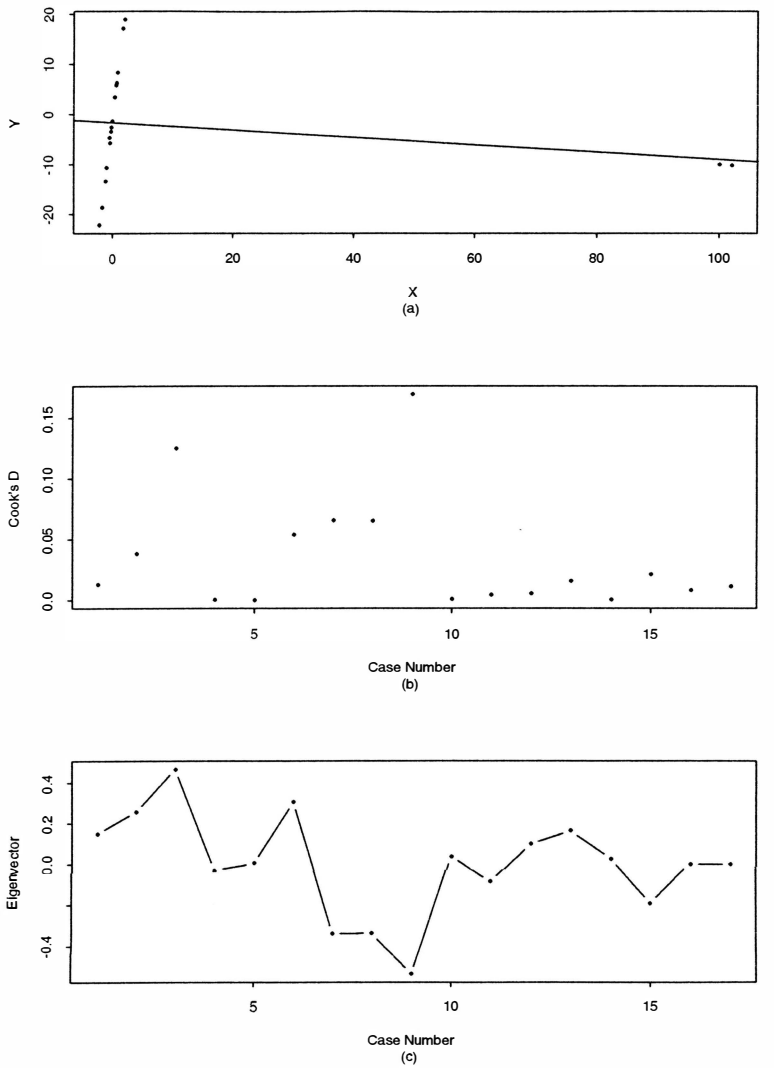

Figure 2.1: An example of masking effect, (a) scatterplot of data, (b) Cook's distance, (c) eigenvector of Cook's approach to local influence

a multivariate point cloud and looking at residuals from a high breakdown fit to identify regression outliers. They combine these two and propose a robust diagnostic plot to identify the outliers and leverage points. However, unlike a local influence approach, their approach shows the influence of individual observations not the joint influence of multiple observations.

The standardized influence matrix approach proposes a measure for the joint influence of a group of cases around the true parameters which will be estimated by robust estimators.

Chapter 3

Standardized Influence Matrix

3.1 Influence Function under Multiple Contamination

The influence function is essentially the Gâteaux differential of a functional (Hampel 1974, 1986). This chapter extends his idea to the case of multiple contamination.

Let Ω be a metric space, let T be a vector-valued mapping from a subset of the probability measures on Ω into the k-dimensional Euclidean space R^k, and let F denote a probability measure on Ω for which T is defined. We consider an estimator as a sequence of statistics $\{T_n : n \geq 1\}$, one for each possible sample size n. Ideally the observations are i.i.d. according to a member of the parametric model $\{F_\theta : \theta \in \Theta\}$, where Θ is an open subset of R^p. And we assume that there exists a functional T such that

$$T_n(x_1, \ldots, x_n) \longrightarrow_{n\to\infty} T(F) \tag{3.1.1}$$

in probability. In addition, we often have asymptotic normality, that is

$$\sqrt{n}[T_n - T(F)] \longrightarrow^{\mathcal{L}}_{n\to\infty} N(0, V(T, F)), \tag{3.1.2}$$

where $V(T, F)$ is called the *asymptotic (variance-)covariance matrix* of $\{T_n; n \geq 1\}$ at F.

Denote by δ_x the probability measure determined by point mass 1 at any given point $x \in \Omega$. Mixtures of F and some δ_x (the contaminated distribution) are written as

$$G = (1 - \epsilon)F + \epsilon\delta_x \quad \text{for } 0 < \epsilon < 1\ . \tag{3.1.3}$$

Then the influence function $IF(\boldsymbol{x}, T, F)$ of T at F is defined pointwise by

$$IF(\boldsymbol{x}, T, F) = \lim_{\epsilon \to 0} \frac{T([(1-\epsilon)F + \epsilon\delta_{\boldsymbol{x}}] - T(F)}{\epsilon} \tag{3.1.4}$$

if this limit is defined for every point $\boldsymbol{x} \in \Omega$.

Since the influence function is a special case of the Gâteaux differential, we have

$$\int IF(\boldsymbol{x}, T, F) dF(\boldsymbol{x}) = 0. \tag{3.1.5}$$

Thus, when G is close to F, the first order von Mises expansion of T at F evaluated in G is given by

$$\begin{aligned} T(G) &= T(F) + \int IF(\boldsymbol{x}, T, F) d(G - F)(\boldsymbol{x}) + \text{remainder} \\ &= T(F) + \epsilon \cdot IF(\boldsymbol{x}, T, F) + o_p(\epsilon). \end{aligned} \tag{3.1.6}$$

where the definition of $o_p(\cdot)$ can be found in Serfling (1980). The asymptotic variance in (3.1.2) equals

$$V(T, F) = \int IF(\boldsymbol{x}, T, F) IF(\boldsymbol{x}, T, F)^T dF(\boldsymbol{x}). \tag{3.1.7}$$

Theorem 3.1 *Suppose that multiple contamination can be expressed as*

$$G = (1-\epsilon)F + \epsilon H, \tag{3.1.8}$$

where $0 < \epsilon < 1$ *and*

$$H = \sum_{i=1}^{q} \pi_i \delta_{x_i} \tag{3.1.9}$$

is a discrete distribution at fixed points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$ *with* $\sum_{i=1}^{q} \pi_i = 1$ *and* $\pi_i \geq 0$ *for* $i = 1, \ldots, q$, *then provided the influence function exists for each* $\boldsymbol{x}_i$, *we have*

$$T(G) - T(F) = \epsilon \cdot E_H[IF(\boldsymbol{x}, T, F)] + o_p(\epsilon) \tag{3.1.10}$$

or

$$T(G) - T(F) = \epsilon \cdot \mathbf{IF}(X, T, F) \cdot \Pi + o_p(\epsilon), \tag{3.1.11}$$

where

$$\mathbf{IF}(X, T, F) = [IF(\boldsymbol{x}_1, T, F), ..., IF(\boldsymbol{x}_q, T, F)], \tag{3.1.12}$$

$X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_q)^T$, *and* $\Pi = (\pi_1, \ldots, \pi_q)^T$.

Proof: By (3.1.6) and (3.1.8), we have

$$\begin{aligned} T(G) &= T(F) + \int IF(\boldsymbol{x}, T, F) d(G - F)(\boldsymbol{x}) + \mathrm{o}_p(\epsilon) \\ &= T(F) + \int IF(\boldsymbol{x}, T, F) \cdot [-\epsilon dF(\boldsymbol{x}) + \epsilon dH(\boldsymbol{x})] + \mathrm{o}_p(\epsilon) \\ &= T(F) + \epsilon \cdot E_H[IF(\boldsymbol{x}, T, F)] + \mathrm{o}_p(\epsilon). \end{aligned}$$

□

In the following, unless it is redefined, $\approx$ will be used to replace equality when $\mathrm{o}_p(\epsilon)$ is omitted.

3.2 Standardized Influence Matrix

To assess the influence of multiple contamination, we need to suggest a certain norm of the influence vector as a measure of influence. Such a measure should be invariant under a strictly monotone transformation on parameters and have a useful interpretation.

Now consider the case of m observations— $X_m = [\boldsymbol{x}_1, ..., \boldsymbol{x}_m]^T$, let $F_{\epsilon,H}$ be a m-point contaminated distribution that can be written as

$$F_{\epsilon,H} = (1 - \epsilon)F + \epsilon H$$

where H is a discrete distribution at X_m. Thus $F_{\epsilon,H}$ allows varying kinds of weight perturbations for all cases through the distribution H. By (3.1.11), we know

$$\begin{aligned} T(F_{\epsilon,H}) - T(F) &\approx \epsilon \cdot E_H[IF(\boldsymbol{x}, T, F)] \\ &= \epsilon \cdot \mathbf{IF}(X, T, F)\Pi. \end{aligned} \tag{3.2.1}$$

Although the class of invariant norms is large the methods of information and self standardization suggested by Hampel et al (1986) are used.

Definition 3.1 *The* **information standardized influence displacement** *is defined as*

$$D_I(H, T(F)) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2}[T(F_{\epsilon,H}) - T(F)]^T \cdot J(T(F)) \cdot [T(F_{\epsilon,H}) - T(F)] \tag{3.2.2}$$

where $J(T(F))$ *is the Fisher information matrix of* $T(F)$.

Obviously, when H is a point mass distribution, $D_I(H, T(F))^{1/2}$ is a norm of the standardized influence function (Ko and Chang 1993), and $\sup_x D_I(\delta_x, T(F))^{1/2}$ is the information standardized sensitivity (Hampel et al 1986). In fact, section 3.5 will show that $D_I(H, T(F))$ is equivalent to likelihood displacement for some T. Now inserting (3.2.1) into (3.2.2), we have

$$D_I(H, T(F)) = \Pi^T \cdot \mathbf{MSIF}_I(X, T, F) \cdot \Pi, \tag{3.2.3}$$

where

$$\mathbf{MSIF}_I(X_m, T, F) = \mathbf{IF}(X, T, F)^T \cdot J(T(F)) \cdot \mathbf{IF}(X, T, F). \tag{3.2.4}$$

Definition 3.2 *The* **information-standardized influence matrix,** $\mathbf{MSIF}_I(X_m, T, F)$ *is given by (3.2.4).*

Similarly, we can define the self-standardized influence displacement.

Definition 3.3 *The* **self-standardized influence displacement** *is defined as*

$$D_S(H, T(F)) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} [T(F_{\epsilon,H}) - T(F)]^T \cdot V(T,F)^{-1} \cdot [T(F_{\epsilon,H}) - T(F)] \qquad (3.2.5)$$

where $V(T,F)^{-1}$ *is the generalized inverse of* $V(T,F)$.

Clearly $D_S(H, T(F))^{1/2}$ is identical to the norm of the self-standardized influence function, and $\sup_x D_S(\delta_x, T(F))^{1/2}$ is the self-standardized sensitivity (Hampel et al 1986). Since in many cases the problem of having singular $V(T,F)$ can be overcome by reparameterizations, we assume $V(T,F)$ is nonsingular in the following. Moreover, section 3.6 will show that it is proportional to the level influence of the test based on the Wald statistic. Now insert (3.2.1) into (3.2.5) and we have

$$D_S(H, T(F)) = \Pi^T \cdot \mathbf{MSIF}_S(X, T, F) \cdot \Pi, \qquad (3.2.6)$$

where

$$\mathbf{MSIF}_S(X_m, T, F) = \mathbf{IF}(X, T, F)^T \cdot V(T,F)^{-1} \cdot \mathbf{IF}(X, T, F). \qquad (3.2.7)$$

Definition 3.4 *The* **self-standardized influence matrix,** $\mathbf{MSIF}_S(X_m, T, F)$ *is given by (3.2.7).*

The ordinary influence function and standardized influence matrix for functional T are most useful in the framework of a parametric model $\{F_\theta; \theta \in \Theta\}$, especially in connection with Fisher-consistent estimators; i.e.,

$$T(F_\theta) = \theta \text{ for all } \theta \in \Theta.$$

Theorem 3.2 *Let* $\eta \in \Theta$, *and suppose that there exists a strictly monotone differentiable mapping* $\xi : \Theta \to \Theta$ *such that* $\xi(\theta) = \eta$. *Now let* U *be the functional giving the parameter value* η; *that is,*

$$U(F_\eta) = \eta.$$

Then

$$\mathbf{MSIF}(X, U, F_\eta) = \mathbf{MSIF}(X, T, F_\theta).$$

Proof: By the definition of the influence function, we have

$$\begin{aligned} IF(\boldsymbol{x}, U, F_\eta) &= \frac{\partial}{\partial \epsilon}[U((1-\epsilon)F_\eta + \epsilon\delta_{\boldsymbol{x}})]_{\epsilon=0} \\ &= \xi'(\theta) \cdot IF(\boldsymbol{x}, T, F_\theta). \end{aligned} \tag{3.2.8}$$

Thus, the information standardized influence matrix

$$\mathbf{MSIF}_I(X, U, F_\eta) \tag{3.2.9}$$

$$= \mathbf{IF}(X, U, F_\eta)^T \cdot J(U(F_\eta)) \cdot \mathbf{IF}(X, U, F_\eta)$$

$$= \mathbf{IF}(X, T, F_\theta)^T \cdot \xi'(\theta)^T \cdot J(U(F_\eta)) \cdot \xi'(\theta) \cdot \mathbf{IF}(X, T, F_\theta). \tag{3.2.10}$$

Since $J(\theta) = \partial^2 L(\theta)/\partial\theta\partial\theta^T$,

$$\begin{aligned} J(U(F_\eta)) = J(\eta) &= \frac{\partial^2 L(\eta)}{\partial\eta\partial\eta^T} \\ &= \xi'^{-1}(\theta)^T \cdot \frac{\partial^2 L(\theta)}{\partial\theta\partial\theta^T} \cdot \xi'^{-1}(\theta) \\ &= \xi'^{-1}(\theta)^T \cdot J(\theta) \cdot \xi'^{-1}(\theta) \\ &= \xi'^{-1}(\theta)^T \cdot J(T(F_\theta)) \cdot \xi'^{-1}(\theta). \end{aligned} \tag{3.2.11}$$

Insert (3.2.11) into (3.2.10) and we have

$$\mathbf{MSIF}_I(X, U, F_\eta) = \mathbf{MSIF}_I(X, T, F_\theta). \tag{3.2.12}$$

Similarly we can show invariance of the self-standardized influence matrix since

$$V(U, F) = \xi'(\theta) V(T, F) \xi'(\theta)^T. \tag{3.2.13}$$

□

3.3 The Local Influence and Standardized Influence Matrix

As we have seen, $D(H, T(F))$, an invariant norm of the vector of the influence function, measures the deviation of parameter estimates from the true parameter by adding X with weight Π. Now for a given X and F, one may think

$$\max_{\Pi} D(H, T(F)) \tag{3.3.1}$$

could serve as a measure of the worst joint influence. However, the solution to (3.3.1) is equivalent to the worst influence by an *individual* observation.

Clearly, the influence by a single observation $\boldsymbol{x}$, as a special case of $D(H, T(F))$, is

$$D(\delta_{\boldsymbol{x}}, T(F)) = IF(\boldsymbol{x}, T, F)^T V(T, F)^{-1} IF(\boldsymbol{x}, T, F). \tag{3.3.2}$$

Now let

$$s_{ij} = IF(\boldsymbol{x}_i, T, F)^T V(T, F)^{-1} IF(\boldsymbol{x}_j, T, F),$$

and we have

$$s_{ii} = D(\delta_{x_i}, T(F)).$$

Moreover, we can rewrite **MSIF** as

$$\mathbf{MSIF}(X_m, T, F) = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \vdots & \vdots \\ s_{m1} & \cdots & s_{mm} \end{pmatrix}. \tag{3.3.3}$$

Theorem 3.3 *For a given X_m and F,*

$$\max_{\Pi} D(H, T(F)) = \max_{i} s_{ii}. \tag{3.3.4}$$

Proof: First of all, by the Cauchy-Schwartz inequality, we have

$$|s_{ij}| \leq \sqrt{s_{ii}}\sqrt{s_{jj}}.$$

Hence, for all $1 \leq i, j \leq m$,

$$|s_{ij}| \leq \max_i s_{ii} = s_{max}.$$

Since $\sum_{i=1}^{m} \pi_i = 1$,

$$\begin{aligned} D(H, T(F)) &= \Pi^T \mathbf{MSIF}(X_m, T, F)\Pi \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}\pi_i\pi_j \\ &\leq \sum_{i=1}^{m}\sum_{j=1}^{m} |s_{ij}|\pi_i\pi_j \\ &\leq \sum_{i=1}^{m}\sum_{j=1}^{m} s_{max}\pi_i\pi_j \\ &= s_{max} \end{aligned}$$

Hence, $D(H, T(F)) \leq s_{max}$.

On the other hand, without loss of generality we can assume $s_{11} = s_{max}$ and we have

$$D(H^*, T(F)) = s_{11}$$

when $\Pi^* = (1, 0, \ldots, 0)^T$. Therefore, we conclude

$$\max_{\Pi} D(H, T(F)) = \max_i s_{ii}.$$

□

The theorem tells us that examination of the maximum value of $D(H, T(F))$ alone is nothing but investigation of the worst influence of single observations, or standardized gross-error sensitivity. Therefore, to study the joint influence by X_m, a new way of examining $D(H, T(F))$ is necessary.

Given the definition of $D(H, T(F))$, it is clear that for a small ϵ, $\epsilon D(H, T(F))^{1/2}$ is a norm of the standardized bias caused by adding X_m with weight Π. Since the rate of

bias with respect to ϵ indicates the worst influence by individual observations, the absolute value of $D(H, T(F))$ does not tell us how sensitive the parameter estimates are in terms of joint influence. However, because the weights represent the inter-relationship among $x_1, \ldots, x_m$, the rate of bias with respect to the real valued weight vector Π could measure the joint influence by X_m as a whole, and the direction that causes the largest changes in bias reveals which alterations simultaneously imposed on X_m may result in considerable changes in parameter estimates. Geometrically, we can inspect the surface defined below:

Definition 3.5 *The* **influence graph** *is defined as the value of the* $(m+1) \times 1$ *vector*

$$\alpha(\Pi) = \begin{pmatrix} \Pi \\ \\ \frac{1}{m} D(H, T(F)) \end{pmatrix} \tag{3.3.5}$$

as Π *varies throughout some open subset of* R^m.

Obviously, the smoother the surface, the less change in bias when a certain action, such as adding or deleting a few points, has been imposed on the observations. However, unless $m \leq 2$, it is impossible for us to visualize an influence graph or exhaustively examine the directional derivatives for all the points on the surface. We need to propose a measure of the behavior of the surface. Thus, we follow the differential geometric method suggested by Cook (1986) and use the normal curvature of the influence graph, which is the curvature of the curve in the cross section of the surface at a direction l. The reason to use curvature instead of slope is that slope of $D(H, T(F))$ at all directions is 0 since it is a quadratic form of Π.

Definition 3.6 *The* **local influence** *is defined as the shape of the influence graph at local point* $\Pi_0 = (\frac{1}{m}, \ldots, \frac{1}{m})^T$, *which is comprised of the maximum curvature of the curve and the corresponding direction that achieves the maximum.*

The following theorem shows that the standardized influence matrix characterizes the local influence.

Theorem 3.4 *The maximum normal curvature is $2/m$ times the largest eigenvalue of the standardized influence matrix; the corresponding eigenvector gives the direction achieving this maximum.*

Proof: Following Cook's geometric normal curvature approach, we construct a normal section by considering a straight line passing through one fixed point $\Pi_0 \in R^m$. Such a line can be represented by

$$\Pi(a) = \Pi_0 + al \tag{3.3.6}$$

where $a \in R^1$ and l is a unit vector in R^m. Thus the normal curvature $C_l = |\ddot{D}(\Pi(a))|$. And it can be evaluated as

$$C_l = \frac{2}{m}|l^T \cdot \mathbf{MSIF}(X, T, F) \cdot l|. \tag{3.3.7}$$

Hence, by matrix theory,

$$C_{max} = \frac{2}{m}\lambda_1,$$

where λ_1 is the largest eigenvalue of **MSIF**, and l_{max} is the corresponding eigenvector.

□

Proposition 3.1 *If the influence function exists for every $x \in \Omega$, the local influence (see definition 3.6) is independent of the choices of Π_0.*

Obviously, the extreme C_{max} denotes a measure of the worst local influence by X. And the l_{max} associated with C_{max} indicates how contamination obtains its maximum local influence. A group of relatively large positive/negative elements of l_{max} indicates that they are relatively influential.

This can be illustrated by application to a location problem with known σ, where $p = 1$ and $F_\theta(x) = F(x - \theta)$. we assume that $\int x dF(x) = \theta$, $F^{-1}(1/2) = \theta$ and $\text{var}(x) = \sigma^2$. Let us start by considering the arithmetic mean of a n-sample. Here $T_n = (1/n)\sum_{i=1}^n x_i$, and the corresponding functional is defined as

$$T_{mean}(F) = \int x dF(x) = \theta.$$

Clearly, $IF(x, T_{mean}, F) = x - \theta$ and $V(T, F) = \sigma^2$. Hence

$$\mathbf{MSIF}(X, T_{mean}, F) = \frac{1}{\sigma^2} \begin{pmatrix} IF(x_1, T_{mean}, F) \\ \vdots \\ IF(x_n, T_{mean}, F) \end{pmatrix} \cdot (IF(x_1, T_{mean}, F), \ldots, IF(x_n, T_{mean}, F)).$$

It can be shown that the (largest) eigenvalue of $\mathbf{MSIF}(X, T_{mean}, F)$ is

$$\lambda_1 = \sum_{i=1}^n (x_i - \theta)^2/\sigma^2,$$

while the corresponding eigenvector is

$$l_1 = \frac{1}{\sigma\sqrt{\lambda_1}} \begin{pmatrix} IF(x_1, T_{mean}, F) \\ \vdots \\ IF(x_n, T_{mean}, F) \end{pmatrix} = \frac{1}{\sigma\sqrt{\lambda_1}} \begin{pmatrix} x_1 - \theta \\ \vdots \\ x_n - \theta \end{pmatrix}.$$

Therefore, the farther away x_i is from the θ, the larger the absolute value of the corresponding element of l_1 and the more influential the x_i. On the other hand, the sign of the elements of l_1 indicates which side the data points reside with respect to center point θ.

Now if we have two very large outliers that are symmetric, and the arithmetic mean is equal to the mean estimate without the outliers, then $D(H, T(F))$ at Π_0 (with equal weights) is equal to 0. However, deleting one of the outliers will cause a large change in the mean estimate, thus the measure of the influence should reflect the large influence of these two observations. In fact, the largest eigenvalue, which is very large because it is the sum of squares, shows the influence. Moreover, since the eigenvector has positive and negative elements, $D(H, T(F))$ is small while the influence is large. Looking at the eigenvector,

it is apparent that adding one observation and deleting another leads to maximum local influence.

In the same framework, one can also examine the sample median. The corresponding functional is given by

$$T(F) = F^{-1}(\frac{1}{2}),$$

and it can be shown that

$$IF(x, T_{med}, F) = \frac{\text{sign}(x - \theta)}{2f(\theta)}$$

and $V(T_{med}, F) = (2f(\theta))^2$ if $f = F'$ is symmetric around θ. Hence the standardized influence matrix of the estimator of sample median is

$$\mathbf{MSIF}(X, T_{med}, F) = \begin{pmatrix} \text{sign}(x_1 - \theta) \\ \vdots \\ \text{sign}(x_n - \theta) \end{pmatrix} \cdot (\text{sign}(x_1 - \theta), \ldots, \text{sign}(x_n - \theta)).$$

Clearly,

$$\lambda_1 = n$$

and

$$l_1 = \frac{1}{\sqrt{n}} \begin{pmatrix} \text{sign}(x_1 - \theta) \\ \vdots \\ \text{sign}(x_n - \theta) \end{pmatrix}.$$

Therefore, in the case of a median estimator, no matter where the x_i is, its influence on the estimator is uniformly bounded. No single or group of points appear to be more influential than the other, as shown by the eigenvector.

3.4 The Standardized Influence Matrix and Principle Component

In this section, the standardized influence matrix is interpreted from the viewpoint of principle components. Like section 3.4, we assume Fisher consistency and existence of influence function.

Since $\boldsymbol{x}$ is a random p-vector, so is $IF(\boldsymbol{x}, T, F)$ with mean 0 and (nonsingular) covariance matrix $V(T, F)$. We follow the definition given by Ko and Chang (1993) and let the standardized influence function be

$$SIF(\boldsymbol{x}, T, F) = V^{-1/2} IF(\boldsymbol{x}, T, F),$$

where $V^{-1/2}(V^{-1/2})^T = V(T, F)^{-1}$. Obviously, $SIF(\boldsymbol{x}, T, F)$ is a p-vector with mean 0 and covariance matrix I_p. Hence, if $\boldsymbol{x}$ follows distribution F, the eigenvalues of the covariance matrix of $SIF(\boldsymbol{x}, T, F)$ are

$$\lambda_1 = \ldots = \lambda_p = 1. \tag{3.4.1}$$

Note that $SIF(\boldsymbol{x}, T, F)$ is not unique since $V^{-1/2}$ is not unique. On the other hand, the length of $SIF(\boldsymbol{x}, T, F)$, or $||SIF(\boldsymbol{x}, T, F)||^2$ remains the same no matter what $V^{-1/2}$ one selects.

Now consider a sample of $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$, we have

$$\mathbf{SIF}(X, T, F) = (SIF(\boldsymbol{x}_1, T, F), \ldots, SIF(\boldsymbol{x}_m, T, F))$$

and

$$\mathbf{MSIF}(X, T, F) = \mathbf{SIF}(X, T, F)^T \mathbf{SIF}(X, T, F).$$

Definition 3.7 *The complement form of the standardized influence matrix is defined by the $p \times p$ matrix,*

$$\mathbf{MSIF}^c(X, T, F) = \mathbf{SIF}(X, T, F)\mathbf{SIF}(X, T, F)^T. \tag{3.4.2}$$

By matrix theory, the eigenvalues of **MSIF** and $\mathbf{MSIF}^c$ are identical.

Theorem 3.5 *The sample covariance matrix of $(SIF(\boldsymbol{x}_1, T, F), \ldots, SIF(\boldsymbol{x}_n, T, F))$ is $(1/n)\mathbf{MSIF}^c(X, T, F)$.*

Proof:

$$\begin{aligned}\frac{1}{n}\mathbf{MSIF}^c(X,T,F) &= \frac{1}{n}\mathbf{SIF}(X,T,F)\mathbf{SIF}(X,T,F)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} SIF(\boldsymbol{x}_i,T,F)SIF(\boldsymbol{x}_i,T,F)^T,\end{aligned}$$

which gives the sample covariance matrix. □

Thus the eigenvalues of the sample covariance matrix are $\hat{\lambda}_1,\ldots,\hat{\lambda}_p$, the eigenvalues of $\mathbf{MSIF}^c$. Hence detecting the existence of influential observations is equivalent to testing the hypothesis

$$H_0 : \lambda_1 = \ldots = \lambda_p = 1. \tag{3.4.3}$$

Moreover, assume $IF(\boldsymbol{x}_1,T,F),\ldots,IF(\boldsymbol{x}_n,T,F)$ are a random sample from a normal distribution, the following large sample distribution theorem for the eigenvalue $\lambda_1,\ldots,\lambda_p$ is given by Johnson and Wichern (1988).

Proposition 3.2 *If $IF(\boldsymbol{x}_1,T,F),\ldots,IF(\boldsymbol{x}_m,T,F)$ are independent identical normally distributed, for large m, the $\hat{\lambda}_i$ are independently distributed with mean $\lambda_i = 1$ and variance $2\lambda_i^2/m = 2/m$.*

Therefore, if $IF(\boldsymbol{x},T,F)$ is normally distributed, a large value of the largest eigenvalue of $\mathbf{MSIF}^c$, or equivalently of **MSIF** indicates the existence of influential observations.

Theorem 3.6 *If e_1 is the eigenvector of $\mathbf{MSIF}^c$ and l_1 is the eigenvector of* **MSIF** *associated with the largest eigenvalue λ_1, then*

$$\begin{aligned} l_1 &= \mathbf{IF}(X,T,F)V^{-1/2}e_1/\sqrt{\lambda_1} \\ &= \mathbf{SIF}(X,T,F)^T e_1/\sqrt{\lambda_1}. \end{aligned} \tag{3.4.4}$$

Moreover, though e_1 depends on which $V^{-1/2}$ is selected, l_1 does not.

Proof: Since e_1 is the eigenvector of $\mathbf{MSIF}^c$ associated with λ_1, we have

$$\mathbf{MSIF}(X,T,F)^c e_1 = \mathbf{SIF}(X,T,F)\mathbf{SIF}(X,T,F)^T = \lambda_1 e_1.$$

By premultiplying $\mathbf{SIF}(X,T,F)^T$ on both sides, we have

$$\mathbf{SIF}(X,T,F)^T\mathbf{SIF}(X,T,F)\mathbf{SIF}(X,T,F)^T e_1 = \lambda_1 \mathbf{SIF}(X,T,F)^T e_1,$$

or

$$\mathbf{MSIF}(X,T,F)l_1' = \lambda_1 l_1'$$

where $l_1' = \mathbf{SIF}(X,T,F)^T e_1$.

On the other hand,

$$\begin{aligned} ||l_1'|| &= (l_1'^T l_1')^{1/2} \\ &= (e_1^T \mathbf{SIF}(X,T,F)\mathbf{SIF}(X,T,F)^T e_1)^{1/2} \\ &= (e_1^T \lambda_1 e_1)^{1/2} \\ &= \sqrt{\lambda_1}. \end{aligned}$$

Hence

$$\begin{aligned} l_1 &= l_1'/||l_1'|| \\ &= \mathbf{SIF}(X,T,F)^T e_1/\sqrt{\lambda_1} \end{aligned}$$

is the eigenvector of $\mathbf{MSIF}(X,T,F)$ associated with λ_1.

Clearly e_1 depends on what $V^{1/2}$ is selected since $\mathbf{MSIF}^c$ depends on $V^{1/2}$; however, l_1, as an eigenvector of **MSIF** is free from the choice of $V^{1/2}$ since **MSIF** is a function of V not of $V^{1/2}$. □

In fact, the ith element of l_1 can be written as

$$l_{1i} = SIF(\boldsymbol{x}_i, T, F)^T e_1/\sqrt{\lambda_1}.$$

By principle component theory, e_1 is the direction in the parameter space that gives the maximum variance for a linear combination of the standardized influence function of each parameter. Hence l_{1i} is the projection of vector $SIF(\boldsymbol{x}_i, T, F)$ along the direction e_1 and measures the contribution of $\boldsymbol{x}_i$ in the direction of e_1. The larger $|l_{1i}|$, the more $SIF(\boldsymbol{x}_i, T, F)$, and ultimately $\boldsymbol{x}_i$, has contributed to the maximum variance, and the more influential is $\boldsymbol{x}_i$.

Recall the concept of local influence from the last section, we may say that the worst local influence corresponds to the maximum variance of a linear combination of the elements of $SIF(\boldsymbol{x}, T, F)$ that is determined by X_m. Furthermore, other influential points could be identified by evaluating eigenvalues $\lambda_1, \ldots, \lambda_p$ and the corresponding eigenvectors.

3.5 The Information Standardization $D_I(H, T(F))$ and Likelihood Displacement

For a given set of data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the log-likelihood $L(\theta)$ is defined as

$$L(\theta) = \sum_{i=1}^{n} \ln[f(\boldsymbol{x}_i, \theta)]$$

where $f(\boldsymbol{x}, \theta)$ is the density function of the underlying distribution F. When using the functional form, we can rewrite $L(\theta)$ as

$$\begin{aligned} L(T(F)) &= n \int \ln[f(t, T(F))] dF_n(t) \\ &= n \cdot E_{F_n}\{\ln[f(t, T(F))]\}. \end{aligned} \tag{3.5.1}$$

Definition 3.8 *The generalized log-likelihood is defined as*

$$\tilde{l}(T(F)) = E_F\{\ln[f(t, T(F))]\}. \tag{3.5.2}$$

Definition 3.9 *The generalized likelihood displacement is given by*

$$\tilde{LD}(\epsilon, H) = 2[\tilde{l}(T(F)) - \tilde{l}(T(F_{\epsilon,H}))]. \tag{3.5.3}$$

Theorem 3.7 *Suppose that the density function f is continuous and differentiable at third order with respect to parameters, and $\partial^2 f/\partial\theta^2$ is continuous,*

$$\tilde{LD}(\epsilon, H) = \epsilon^2 D_I(H, T(F)) + o_p(\epsilon^2). \tag{3.5.4}$$

Proof: By Taylor's expansion, we can rewrite $f(t, T(F_{\epsilon,H}))$ as

$$\begin{aligned}
&\ln[f(t, T(F_{\epsilon,H}))] \\
&= \ln[f(t, T(F))] + \left.\frac{\partial \ln[f(t, T(F_{\epsilon,H})]}{\partial\epsilon}\right|_{\epsilon=0} \cdot \epsilon \\
&\quad + \frac{1}{2}\epsilon^2 \cdot \left.\frac{\partial^2 \ln[f(t, T(F_{\epsilon,H}))]}{\partial\epsilon^2}\right|_{\epsilon=0} + o_p(\epsilon^2) \\
&= \ln[f(t, T(F))] + \left.\frac{\partial \ln[f(t,\theta)]}{\partial\theta}\right|_{\theta=T(F)} \cdot \epsilon E_H[IF(\boldsymbol{x}, T, F)] \\
&\quad + \frac{1}{2}\epsilon^2 E_H[IF(\boldsymbol{x}, T, F)]^T \cdot \left.\frac{\partial^2 \ln[f(t,\theta)]}{\partial\theta\partial\theta^T}\right|_{\theta=T(F)} \cdot E_H[IF(\boldsymbol{x}, T, F)] \\
&\quad + o_p(\epsilon^2).
\end{aligned} \tag{3.5.5}$$

Then

$$\begin{aligned}
\tilde{LD}(\epsilon, H) &= 2\left\{\int \ln[f(t, T(F))]dF(t) - \int \ln[f(t, T(F_{\epsilon,H}))]dF(t)\right\} \\
&= 2\int \{\ln[f(t, T(F)] - \ln[f(t, T(F_{\epsilon,H})]\}dF(t) \\
&= \int \left\{-2\left.\frac{\partial \ln[f(t,\theta)]}{\partial\theta}\right|_{\theta=T(F)} \cdot \epsilon E_H[IF(\boldsymbol{x}, T, F)]\right. \\
&\quad -\epsilon^2 E_H[IF(\boldsymbol{x}, T, F)]^T \cdot \left.\frac{\partial^2 \ln[f(t,\theta)]}{\partial\theta\partial\theta^T}\right|_{\theta=T(F)} \cdot E_H[IF(\boldsymbol{x}, T, F)] \\
&\quad \left. + o_p(\epsilon^2)\right\} dF(t) \\
&= \left\{-2\int \left.\frac{\partial \ln[f(t,\theta)]}{\partial\theta}\right|_{\theta=T(F)} dF(t) \cdot \epsilon E_H[IF(\boldsymbol{x}, T, F)]\right. \\
&\quad \left. + \epsilon^2 E_H[IF(\boldsymbol{x}, T, F)]^T \cdot \int -\left.\frac{\partial^2 \ln[f(t,\theta)]}{\partial\theta\partial\theta^T}\right|_{\theta=T(F)} dF(t) \cdot E_H[IF(\boldsymbol{x}, T, F)]\right\} \\
&\quad + o_p(\epsilon^2).
\end{aligned}$$

Under assumptions of interchangeability of the derivative and integral (see Rudin 1976), it can be shown that

$$\int \frac{\partial \ln[f(t,\theta)]}{\partial \theta} dF(t) = 0 \tag{3.5.6}$$

and

$$J(T(F)) = \int -\frac{\partial^2 \ln[f(t,\theta)]}{\partial\theta\partial\theta^T}|_{\theta=T(F)} dF(t). \tag{3.5.7}$$

Using (3.5.6) and (3.5.7), we have

$$\begin{aligned} \tilde{LD}(\epsilon, H) &= \epsilon^2 E_H[IF(x,T,F)]^T \cdot J(T(F)) \cdot E_H[IF(x,T,F)] + o_p(\epsilon^2) \\ &= \epsilon^2 D_I(H, T(F)) + o_p(\epsilon^2). \end{aligned}$$

□

Theorem 3.8 *As a result of generalized likelihood displacement, Cook's approach to local influence is equivalent to the information standardized approach when $T(F)$ and $T(F_{\epsilon,H})$ are replaced by the maximum likelihood estimators $\hat{\theta}$ and $\hat{\theta}_\omega$ with $\epsilon = O(1/n)$, $\Pi = (\omega_0 + \frac{\omega-\omega_0}{\epsilon})/n$ and $\omega_0 = (1,\ldots,1)^T$, respectively. Specifically,*

$$LD(\omega) = n\epsilon^2 D_I(H, T(F_n)) + o_p(\epsilon^2) \tag{3.5.8}$$

and the maximum curvature of the influence graph given by Cook (see section 2.2.2) is

$$C_l = 2|l^T \cdot \mathbf{MSIF}_I(X, T, F_n) \cdot l|/n. \tag{3.5.9}$$

Proof: Recall Cook's approach to local influence: he defined the influence graph as (2.2.6) where $LD(\omega)$ is defined as (2.2.5). Since $\hat{\theta}$ and $\hat{\theta}_\omega$ are the maximum likelihood estimators based on $L(\theta)$ and $L(\theta|\omega)$, by Taylor expansion we can rewrite $L(\hat{\theta}_\omega)$ as

$$L(\hat{\theta}_\omega) = L(\hat{\theta}) + L'(\hat{\theta}) \cdot (\hat{\theta}_\omega - \hat{\theta}) + \frac{1}{2}(\hat{\theta}_\omega - \hat{\theta})^T L''(\hat{\theta})(\hat{\theta}_\omega - \hat{\theta})$$

$$
\begin{aligned}
&+o_p(||\hat{\theta}_\omega - \hat{\theta}||^2) \\
\approx\ & L(\hat{\theta}) - \frac{n}{2}(\hat{\theta}_\omega - \hat{\theta})^T J(\hat{\theta})(\hat{\theta}_\omega - \hat{\theta}),
\end{aligned}
$$

thus,

$$
\begin{aligned}
LD(\omega) &= 2[L(\hat{\theta}) - L(\hat{\theta_\omega})] \\
&\approx n(\hat{\theta}_\omega - \hat{\theta})^T J(\hat{\theta})(\hat{\theta}_\omega - \hat{\theta}). \qquad (3.5.10)
\end{aligned}
$$

If we write $\hat{\theta}$ and $\hat{\theta}_\omega$ in functional form, they are $T(F_n)$ and $T(F_{n_{\epsilon,H}})$ respectively, where

$$\Pi = (\omega_0 + \frac{\omega - \omega_0}{\epsilon})/n$$

with $\epsilon = O(1/n)$ and F_n and $F_{n_{\epsilon,H}}$ are empirical distributions of F and $F_{\epsilon,H}$ respectively, then

$$
\begin{aligned}
LD(\omega) &= n[T(F_{n_{\epsilon,H}}) - T(F_n)]^T \cdot J(T(F_n)) \cdot [T(F_{n_{\epsilon,H}}) - T(F_n)] + o_p(\epsilon^2) \\
&= n\epsilon^2 D_I(H, T(F_n)) + o_p(\epsilon^2).
\end{aligned}
$$

This result shows the equivalence of likelihood displacement and information standardized influence displacement when the parameters are approximated by maximum likelihood estimators.

Furthermore, the curvature of Cook's influence graph is

$$C_l = \frac{d^2 LD(\omega(a))}{da^2}$$

where $\omega(a) = \omega_0 + al$. Hence,

$$
\begin{aligned}
C_l &= \frac{d^2}{da^2} n\epsilon^2 \Pi(a)^T \mathbf{MSIF}_I(X, T, F_n)\Pi(a) \\
&= \frac{1}{n}\epsilon^2 \frac{d^2}{da^2}\left[(\omega_0 + \frac{al}{\epsilon})^T \mathbf{MSIF}_I(X, T, F_n)(\omega_0 + \frac{al}{\epsilon})\right] \\
&= \frac{2}{n}|l^T \mathbf{MSIF}_I(X, T, F_n) l|.
\end{aligned}
$$

□

Therefore, the maximum curvature proposed by Cook (1986) is the maximum curvature of the information standardized approach around the maximum likelihood estimator.

3.6 Self-Standardization $D_S(H, T(F))$ and Wald Test Statistic

Although information standardization is useful to assess the local influence of various kinds of estimators, it makes more sense to study self-standardization influence for estimators other than the maximum likelihood estimator. Note that the information standardized matrix can be considered as a special case of the self-standardized influence matrix, in which $T(F)$ represents the maximum likelihood estimator. In the following, the standardized influence matrix refers to the self-standardized influence matrix. And we drop the subscript unless it is needed for clarity.

This section is going to show that the self-standardized influence matrix characterizes the local influence on the level of the Wald test for the hypothesis $H_0 : \theta = \theta_0$.

Now suppose we want to test the hypothesis $H_0 : \theta = \theta_0$ based on a consistent estimator $\hat{\theta}$, the Wald test statistic is given by

$$W_n^2 = (\hat{\theta} - \theta_0)^T V^{-1} (\hat{\theta} - \theta_0) \tag{3.6.1}$$

where V is the (nonsingular) variance-covariance matrix of $\hat{\theta}$ and is independent of $\hat{\theta}$. We define $T(F)$ and $W(F)$ to be the functional corresponding to $\hat{\theta}$ and W_n such that

$$T(F_\theta) = \theta,$$

$$W(F) = \{(T(F) - \theta_0)^T V(T, F)^{-1} (T(F) - \theta_0)\}^{1/2}, \tag{3.6.2}$$

where we assume $V(T, F)$ is known, since in many circumstances it is only a function of a

few scale parameters. Hence, we have

$$W_n^2 = nW^2(F_n)$$

and we can reduce the null hypothesis to

$$H_0 : W^2(F) = 0.$$

In order to study the asymptotic power of the test, one usually constructs a sequence of alternatives $\theta_n = \theta_0 + \Delta n^{-1/2}$. Similarly, the alternative hypotheses can be reduced to

$$H_n : W^2(F) = \frac{1}{n}\Delta^T V(T,F)^{-1}\Delta.$$

Definition 3.10 *The power influence displacement of ϵ-contamination for a test based on Wald statistics is given by*

$$PD_\theta(\epsilon, H, W^2(F)) = F_{\theta_\epsilon, H}(W_n^2 \geq C_{p,\alpha}) - F_\theta(W_n^2 \geq C_{p,\alpha}), \tag{3.6.3}$$

where $C_{p,\alpha}$ is the $1-\alpha$ quantile of the Chi-square distribution with p degrees of freedom.

Note that the level influence displacement is a special case of the power influence displacement where $\theta = \theta_0$.

Theorem 3.9 *The level influence displacement can be written as*

$$PD_{\theta_0}(\epsilon, H, W^2(F)) = d_0 \cdot n\epsilon^2 D_S(H, T(F)) + o_p(\epsilon^2) \tag{3.6.4}$$

where d_0 is a constant that is given by

$$d_0 = \int_{c_{p,\alpha}}^{\infty} \left.\frac{\partial f(u,w)}{\partial w}\right|_{w=0} du,$$

and $f(u,w)$ is the density function of the Chi-square distribution with p degrees of freedom and noncentrality parameter w.

Proof: Consider a n-point contamination H,

$$W^2(F_{\epsilon,H}) - W^2(F) = (T(F_{\epsilon,H}) - \theta_0)^T V(T,F)^{-1}(T(F_{\epsilon,H}) - \theta_0)$$
$$-(T(F) - \theta_0)^T V(T,F)^{-1}(T(F) - \theta_0). \qquad (3.6.5)$$

Using (3.1.11), we have

$$\begin{aligned} & W^2(F_{\epsilon,H}) - W^2(F) \\ &= (T(F) + \epsilon E_H[IF(\boldsymbol{x},T,F)] - \theta_0)^T V(T,F)^{-1}(T(F) + \epsilon E_H[IF(\boldsymbol{x},T,F)] - \theta_0) \\ &\quad -(T(F) - \theta_0)^T V(T,F)^{-1}(T(F) - \theta_0) + o_p(\epsilon^2) \\ &= 2\epsilon(T(F) - \theta_0)^T V(T,F)^{-1} E_H[IF(\boldsymbol{x},T,F)] \\ &\quad +\epsilon^2 E_H[IF(\boldsymbol{x},T,F)]^T V(T,F)^{-1} E_H[IF(\boldsymbol{x},T,F)] + o_p(\epsilon^2). \end{aligned}$$

Under the null hypothesis $H_0 : T(F) = \theta_0$,

$$W^2(F_{\epsilon,H}) - W^2(F) = \epsilon^2 D_S(H, T(F)) + o_p(\epsilon^2). \qquad (3.6.6)$$

Under the alternative hypotheses $H_n : T(F) = \theta_0 + \Delta n^{-1/2}$,

$$\begin{aligned} & W^2(F_{\epsilon,H}) - W^2(F) \\ &= \frac{2\epsilon}{\sqrt{n}} \Delta^T V(T,F)^{-1} E_H[IF(X,T,F)] \\ &\quad +\epsilon^2 E_H[IF(\boldsymbol{x},T,F)]^T V(T,F)^{-1} E_H[IF(\boldsymbol{x},T,F)] + o_p(\epsilon^2), \end{aligned} \qquad (3.6.7)$$

or

$$\mathbf{IF}(X, W^2, F_{\theta_n}) = \frac{2}{\sqrt{n}} \Delta^T V(T,F)^{-1} \mathbf{IF}(X,T,F). \qquad (3.6.8)$$

Now we want to find the asymptotic distribution of W_n^2. Since $\hat{\theta} \sim AN(\theta, V)$, $W_n^2 \sim \chi_p^2(\delta)$, where p is the rank of V and δ is the noncentrality parameter given by

$$\delta = n(\theta - \theta_0)^T V(T,F)^{-1}(\theta - \theta_0). \qquad (3.6.9)$$

Now let $f(u,\delta)$ be the density function of a random variable u with $\chi^2_{p,\delta}$, then the critical values $C_{p,\alpha}$ of W_n^2 are given by

$$F_{\theta_0}(W_n^2 \geq C_{p,\alpha}) = \alpha,$$

and the power function is defined as

$$\begin{aligned} P_F(\theta) &= F_\theta(W_n^2 \geq C_{p,\alpha}) && (3.6.10)\\ &= \int_{C_{p,\alpha}}^{\infty} f(u,\delta)du. && (3.6.11) \end{aligned}$$

Returning to the n-point contamination H, we have $\hat{\theta} \sim AN(T(F_{\epsilon,H}), V)$ and

$$W_n^2 \sim \chi_p^2(\delta_{\epsilon,H}), \tag{3.6.12}$$

where

$$\begin{aligned} \delta_{\epsilon,H} &= n(\theta_{\epsilon,H} - \theta_0)^T V(T,F)^{-1}(\theta_{\epsilon,H} - \theta_0) && (3.6.13)\\ &= \delta + 2n\epsilon(\theta - \theta_0)^T V(T,F)^{-1} E_H[IF(x,T,F)] \\ &\quad + n\epsilon^2 E_H[IF(x,T,F)]^T V(T,F)^{-1} E_H[IF(x,T,F)] + o_p(\epsilon^2). && (3.6.14) \end{aligned}$$

Then

$$\begin{aligned} PD(\epsilon, H, W^2(F)) &= F_{\theta_{\epsilon,H}}(W_n^2 \geq C_{p,\alpha}) - F_\theta(W_n^2 \geq C_{p,\alpha}) && (3.6.15)\\ &= \int_{C_{p,\alpha}}^{\infty} [f(u,\delta_{\epsilon,H}) - f(u,\delta)]du && (3.6.16)\\ &= \int_{c_{p,\alpha}}^{\infty} \left.\frac{\partial f(u,w)}{\partial w}\right|_{w=\delta} du \cdot (\delta_{\epsilon,H} - \delta) + o_p(\epsilon^2). && (3.6.17) \end{aligned}$$

The level influence displacement is

$$PD_{\theta_0}(\epsilon, H, W^2(F)) = d_0 \cdot n\epsilon^2 D_S(H, T(F)) + o_p(\epsilon^2).$$

□

As a result of this proof, the power influence displacement of the test can be written as

$$\begin{aligned} PD_{\theta_n}(\epsilon, H, W^2) &= d_\Delta \cdot \{2\sqrt{n}\epsilon\Delta^T V(T,F)^{-1} E_H[IF(\boldsymbol{x},T,F)] \\ &\quad + n\epsilon^2 E_H[IF(\boldsymbol{x},T,F)]^T V(T,F)^{-1} E_H[IF(\boldsymbol{x},T,F)]\} \\ &\quad + o_p(\epsilon^2), \end{aligned} \tag{3.6.18}$$

where d_Δ is a constant for a given Δ or θ_n, which is given by

$$d_\Delta = \int_{c_{p,\alpha}}^{\infty} \left.\frac{\partial f(u,w)}{\partial w}\right|_{w=\delta_n} du.$$

Therefore, we can conclude that the level influence displacement of the Wald test is proportional to the self-standardized influence displacement of the corresponding parameter estimators, and the local influence on the level of the Wald test is equivalent to the local influence on parameter estimates. However, the power influence displacement of Wald test is first order in ϵ, which depends on the influence function of parameter estimators as well as the alternative hypotheses. In other words, more deterioration occurs in the power function than in the level of the test for a given rate of contamination.

Chapter 4

Robustness Measures and Diagnostics

As discussed in chapter 1, there are usually two approaches dealing with effects of contaminations. One approach, robust estimators, is to accommodate the gross error that data have. The other one, diagnostic procedures, is to reject outliers according to some criteria. In fact, each approach complements the other. The rejection approach, in some sense, is a non-continuous robust procedure, while a robust estimator often triggers an effective diagnostic method.

As we have seen in the previous chapter, the standardized influence matrix characterizes the local influence on inference of data. A new robustness measure based on the standardized influence matrix could serve as a supplementary concept to existing robustness measures. Moreover, the standardized influence matrix aided by certain robust estimators becomes an effective diagnostic tool in detecting influential observations.

4.1 Existing Robustness Measures

The most important robustness requirements are a low *gross-error sensitivity*, a high *breakdown point* and *qualitative robustness*.

4.1.1 Measures Based on Influence Function

From the robustness point of view, there are at least three important summary values of the influence function apart from its expected values. They were introduced by Hampel (1968, 1974, 1986). The first and most important one is the supremum of a norm of the influence function, *gross-error sensitivity* of T at F. It is defined as

$$\gamma^*(T,F) = \sup_x |IF(\boldsymbol{x},T,F)|. \tag{4.1.1}$$

where the supremum being taken over all $\boldsymbol{x}$ where $IF(\boldsymbol{x},T,F)$ exists. The gross-error sensitivity measures the worst (approximate) influence which a small amount of contamination of fixed size can have on the value of the estimator. Therefore, it may be regarded as an upper bound on the asymptotic bias of the estimator. It is a desirable feature that $\gamma^*(T,F)$ be finite, in which case T is called *B-robust* at F.

Since γ^* is not an invariant measure, Krasker and Welsch (1982) defined an invariant sensitivity measure

$$\gamma_S^{**} = \sup_x [IF(\boldsymbol{x},T,F)^T V(T,F)^{-1} IF(\boldsymbol{x},T,F)]^{1/2}, \tag{4.1.2}$$

which we call *self-standardized gross-error sensitivity*. Similarly we can define the *information-standardized gross-error sensitivity* as

$$\gamma_I^{**} = \sup_x [IF(\boldsymbol{x},T,F)^T J(T(F)) IF(\boldsymbol{x},T,F)]^{1/2}. \tag{4.1.3}$$

The second summary value has to do with small fluctuations in the observations. A measure for the worst effect of "wiggling" (slightly changing a value) is provided by *local-shift sensitivity*

$$\lambda^* = \sup_{x \neq y} \frac{|IF(y,T,F) - IF(X,T,F)|}{|y - \boldsymbol{x}|}. \tag{4.1.4}$$

The third summary value, the *rejection point* is defined by

$$\rho^* = \inf\{r > 0; IF(\boldsymbol{x}, T, F) = 0 \text{ when} |\boldsymbol{x}| > r\}. \tag{4.1.5}$$

(If there exists no such r, then $\rho^* = \infty$ by definition of the infimum.) All observations farther away than ρ^* are rejected completely. Therefore, it is a desirable feature if ρ^* is finite.

4.1.2 Global Reliability: The Breakdown Point

Since the influence function is a collection of directional derivatives in the directions of the mass point δ_x, and is evaluated at the distribution described by the model, it is a *local* concept. Hence it must be complemented by a measure of *global* reliability of the estimator. The concept of a finite sample breakdown point was first suggested by Hodges (1967) as "tolerance of extreme values" in the context of the location problem. It was generalized to a statistical functional by Hampel (1968). However the general formulation of a breakdown point was asymptotic and rather mathematical in nature. In this section we present a simple finite-sample version of the breakdown point, introduced by Donoho and Huber (1983).

Roughly speaking, the breakdown point is the smallest fraction of data contamination needed to cause an arbitrarily large change in the estimate. Let $X_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be any sample of size n and let $T = \{T_n\}_{n \geq p}$ be a sequence of estimates of θ. Now consider all possible corrupted samples $X'_{n,m}$ that are obtained by replacing any m of the original data points by arbitrary values. Define the bias$(m; T, X_n)$ as the maximum bias that can be caused by such a contamination:

$$\text{bias}(m; T, X_n) = \sup_{X'_{n,m}} |T(X'_{n,m}) - T(X_n)|, \tag{4.1.6}$$

where the supremum is over all possible $X'_{n,m}$. Then the breakdown point of T at the sample

X_n is defined as

$$\epsilon_n^*(T, X_n) = \min\left\{\frac{m}{n}; \text{bias}(m; T, X_n) = \infty\right\}. \tag{4.1.7}$$

4.1.3 Qualitative Robustness

Hampel (1971) also introduced some qualitative notions. The main idea is to complement the notion of differentiability (which leads to the influence function) with continuity conditions with respect to the Prohorov distance. Since it yields only a simple dichotomy and is hardly used in the dissertation, this section only gives the definitions for the Prohorov distance and qualitative robustness.

The *Prohorov distance* (Prohorov 1956) of two probability distributions F and G is given by

$$\pi(F, G) = \inf\{\epsilon; F(A) \leq G(A^\epsilon) + \epsilon \text{ for all events } A\} \tag{4.1.8}$$

where A^ϵ is the set of all points whose distance from A is less than ϵ.

We say that a sequence of estimators $\{T_n; n \geq 1\}$ is *qualitatively robust* at F if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all G and for all n:

$$\pi(F, G) < \delta \Rightarrow \pi(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \epsilon \tag{4.1.9}$$

where $\mathcal{L}_F(T_n)$ means the distribution of T_n under F.

4.2 Joint Local Influence

To introduce the concept of bounded joint local influence, we define

$$s_{ij} = IF(x_i, T, F)^T V(T, F)^{-1} IF(x_j, T, F), \tag{4.2.1}$$

for all $i, j = 1, \ldots, m$.

Definition 4.1 *The single local influence of* $T(F)$ *at a point* x_i *is defined as* s_{ii}.

Then bounded gross-error sensitivity is equivalent to bounded single local influence, i.e., $s_{ii} \leq a^2$ for $i = 1, \ldots, m$ (a is a specified constant). On the other hand, the standardized influence matrix can be rewritten as

$$\mathbf{MSIF}(X, T, F) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}. \tag{4.2.2}$$

Definition 4.2 *The joint local influence of* $T(F)$ *at* $X = (x_1, \ldots, x_q)^T$ *is defined as* $\lambda_m(X)$, *the largest eigenvalue of* **MSIF**.

Since **MSIF** is symmetric and semi-positive definite, we have

$$\frac{1}{p}\sum_{i=1}^{m} s_{ii} \leq \lambda_m(X) \leq \sum_{i=1}^{m} s_{ii}, \tag{4.2.3}$$

where p is the dimension of the parameter.

Proposition 4.1 *For a given* m, *bounded joint local influence is equivalent to bounded single local influence (gross-error sensitivity).*

However, since λ_m increases as m increases, eventually bounded λ_m will be impossible for all m. Even when $s_{ii} < a^2$ for all $i = 1, \ldots, m$, (4.2.3) may lead to

$$\sup_X \lambda_m(X) \to \infty \text{ as } m \to \infty.$$

Now suppose $IF(x, T, F)$ is normally distributed, by proposition 3.2, we have the following asymptotic property:

$$\frac{\lambda_m(X)}{m} \sim AN(1, 2/m),$$

or

$$\sqrt{m}\left[\frac{\lambda_m(X)}{m} - 1\right] \sim AN(0, 2).$$

Definition 4.3 *The estimator $T(F)$ has bounded joint local influence if its joint local influence $\lambda_m(X)$ satisfies*

$$\sqrt{m}\sup_X \left|\frac{\lambda_m(X)}{m} - 1\right| \leq C, \ \ as\ m \to \infty \tag{4.2.4}$$

where C is a constant.

Proposition 4.2 *If estimator $T(F)$ has bounded joint local influence, then*

$$\frac{\lambda_m(X)}{m} \to 1, \ \ as\ m \to \infty.$$

Proposition 4.3 *Suppose we have a sample of $\{x_1, \ldots, x_n\}$, the joint local influence by $\{x_1, \ldots, x_m\}$ where $m < n$, and*

$$\lim_{m,n\to\infty} \frac{m}{n} = 0,$$

has the following property:

$$\lim_{m,n\to\infty} \frac{\lambda_m(X)}{n} = 0.$$

Therefore, the estimator with bounded joint local influence and $\lim m/n = 0$ will lead to ignorable influence. On the other hand, it is easy to show that an estimator with unbounded standardized gross-error sensitivity has unbounded $\sup_X \lambda_m(X)/m$, and consequently has unbounded joint local influence.

Proposition 4.4 *If an estimator $T(F)$ has unbounded single local influence (standardized gross-error sensitivity), then it has unbounded joint local influence.*

Besides the concept of bounded joint influence, the standardized influence matrix provides a tool to measure the robustness of different estimators in terms of the joint local influence. In particular, given a dataset, one can calculate the joint influence λ_n for several kinds of estimators. Obviously, the larger λ_n is, the less robust is the estimator, at least

for specific data. The next few chapters formulate the standardized influence matrix for several proposed estimators in linear and logistic regression, and examine their robustness in relation to specific datasets.

4.3 Joint Local Influence and Breakdown Point

Recall from chapter 3 that a contaminated distribution is expressed as

$$F_{\epsilon,H} = (1-\epsilon)F + \epsilon H, \tag{4.3.1}$$

and the bias function is defined by

$$b(\epsilon, H, F) = T(F_{\epsilon,H}) - T(F). \tag{4.3.2}$$

In fact, both joint local influence and the breakdown point of the estimator contribute to the bias, but in different ways.

- The breakdown point of the estimator,

$$\epsilon^* = \min\{\epsilon : \sup_H |b(\epsilon, H, F)| = \infty\}, \tag{4.3.3}$$

  is the minimum of the ratio of the contamination that will make the estimator unacceptable if all kinds of contamination are considered. As pointed out in the previous section, it is a global property.

- Joint local influence represents the slope of bias at the true distribution F with the contamination H_m, where $m/n \to 0$.

To further understand the relationship between the joint local influence and breakdown point, we introduce a sample version of the bias function

$$b(m, H_m, F_{n-m}) = T[(1-\frac{m}{n})F_{n-m} + \frac{m}{n}H_m] - T(F_{n-m}). \tag{4.3.4}$$

Then

$$\epsilon_n^* = \min\{\frac{m}{n} : |b(m, H_m, F_{n-m})| = \infty\}. \tag{4.3.5}$$

For small m/n, in the sense that $\frac{m}{n} \to 0$ as $n \to \infty$,

$$b(m, H_m, F_{n-m}) \approx \frac{m}{n} E_{H_m}[IF(\boldsymbol{x}, T, F_{n-m})]. \tag{4.3.6}$$

Therefore, boundedness of the influence function is equivalent to a nonzero breakdown point.

- If the influence function of T at F is unbounded,

  $$\epsilon_n^* = \frac{1}{n}.$$

  Hence, $\epsilon^* = 0$.

- If the influence function of T at F is bounded,

  $$\epsilon_n^* > \frac{m}{n}.$$

  Hence it has nonzero breakdown point.

Nevertheless, neither single local influence nor the joint local influence can help to determine the breakdown point of an estimator.

4.4 Diagnostics

An alternative approach to dealing with influential points is to construct diagnostic statistics. Chapter 2 has reviewed various kinds of diagnostics that focus attention on observations having a large influence on least squares or maximum likelihood estimators. The diagnostic procedure we are presenting provides statistics that reveal the observations having a large joint influence not only on least squares or maximum likelihood estimators but on any estimators that can be expressed as a functional.

Recall from chapter 3, the standardized influence matrix characterizes the influence graph, the largest eigenvalue of the matrix represents the maximum curvature of the graph evaluated at true parameters, and the corresponding eigenvector indicates the direction of contamination that obtains the maximum influence (maximum curvature). Hence, the eigenvalue of the standardized influence matrix shows the severity of joint local influence on the inference of parameter estimators, while the associated eigenvector identifies the observations that are most influential.

However, there are two problems with the effectiveness of diagnostics. First, the standardized influence matrix is a function of the sample X as well as the unknown true parameters. To obtain the diagnostics, one has to replace the true parameter with a certain robust estimator so that they won't be too much influenced by the data points. The eligible robust estimators will be discussed in the next few chapters.

The second problem is calibration of the largest eigenvalue, which may depend on which estimator we are studying. The following provides a general but not accurate guideline, which we call the np-guideline later.

Theorem 4.1

$$E\,trace[\mathbf{MSIF}(X,T,F)] = np$$

where n is the sample size of X and p is the number of parameters.

Proof:

$$\begin{aligned}
\text{trace}(\mathbf{MSIF}) &= \text{trace}(\mathbf{IF}(X,T,F)^T V(T,F)^{-1}\mathbf{IF}(X,T,F)) \\
&= \text{trace}(\mathbf{IF}(X,T,F)\mathbf{IF}(X,T,F)^T V(T,F)^{-1}) \\
&= \text{trace}(\sum_{i=1}^{n} IF(\boldsymbol{x}_i,T,F)IF(\boldsymbol{x}_i,T,F)^T V(T,F)^{-1})
\end{aligned}$$

Then if all $\boldsymbol{x}_i$ are the sample from distribution F, we have

$$E[\sum_{i=1}^{n} IF(\boldsymbol{x}_i, T, F) IF(\boldsymbol{x}_i, T, F)^T] = nV(T, F).$$

Thus,

$$\begin{aligned} E[\text{trace}(\mathbf{MSIF})] &= \text{trace}(nV(T, F)V(T, F)^{-1}) \\ &= np \end{aligned}$$

□

Since the standardized influence matrix **MSIF** is a semi-positive definite symmetric matrix with rank p, we have

$$\text{trace}(\mathbf{MSIF}) = \sum_{k=1}^{p} \lambda_k,$$

where $\lambda_1 \geq \ldots \geq \lambda_p$. Therefore, if the largest eigenvalue is larger than np, then we say the joint local influence by X is severe and there are joint influential points in X, which can be identified by looking at the corresponding eigenvector. But if the largest eigenvalue is less than np, a more accurate reference depends on the specific estimator. For more details, see the next few chapters.

Chapter 5

Linear Regression Applications

In this chapter the standardized influence matrix is applied in a linear regression context. We start with developing the standardized influence matrix for a least squares estimator, M-estimator and bounded influence estimator. Then global robust estimators are reviewed to construct a joint influence measure for the estimators. Examples to illustrate merits of this technique will be found in chapter 9.

We consider the following linear model. Let $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ be a sequence of independent identically distributed random variables such that

$$y_i = \boldsymbol{x}_i\theta + \epsilon_i, \ i = 1, \ldots, n, \tag{5.0.1}$$

where $y_i \in \mathbf{R}$ is the ith observation, $\boldsymbol{x}_i \in \mathbf{R}^p$ is the ith row of the design matrix $X_{n\times p}$, $\theta \in \Theta \subset \mathbf{R}^p$ is a p-vector of unknown parameters ($p \geq 1$) and $\epsilon_i \in \mathbf{R}$ is the ith error.

We assume that Θ is open and convex and that ϵ_i is independent of $\boldsymbol{x}_i$ and has a symmetric distribution $G(\cdot/\sigma)$, where $\sigma > 0$ is a scale parameter, with density g with respect to the Lebesgue measure. Finally we assume $\boldsymbol{x}_i, i = 2, \ldots, n$ (excluding an intercept term) has mean 0 and finite second moment.

5.1 Least Squares Estimator

The well known least squares (LS) estimator is defined as

$$\min_{\theta} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i\theta)^2, \tag{5.1.1}$$

or is the solution to

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i\theta)\boldsymbol{x}_i^T = 0. \tag{5.1.2}$$

The functional $T(F)$ corresponding to the LS estimator is the solution of

$$\int (y - \boldsymbol{x} \cdot T(F))\boldsymbol{x}^T dF(\boldsymbol{x}, y) = 0, \tag{5.1.3}$$

where F(x,y) is the joint distribution of $(\boldsymbol{x}, y)$, which is assumed to have a density

$$f_\theta(\boldsymbol{x}, y) = g((y - \boldsymbol{x}\theta)/\sigma)k(\boldsymbol{x}). \tag{5.1.4}$$

It can be shown that the influence function of $T(F)$ is

$$IF(\boldsymbol{x}, T, F) = \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1}(y - \boldsymbol{x} \cdot T(F))\boldsymbol{x}^T \tag{5.1.5}$$

and

$$V(T, F) = \sigma^2\{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1}. \tag{5.1.6}$$

Thus,

$$\mathbf{MSIF}(X, Y; T, F) = RX\{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1}X^T R/\sigma^2 \tag{5.1.7}$$

where $R = \text{diag}(r_1, \ldots, r_n)$ and $r_i = y_i - \boldsymbol{x}_i\theta$ for $i = 1, \ldots, n$. As is well known, the LS estimator has unbounded gross-error sensitivity or single local influence. Based on the argument in chapter 3, it has unbounded joint local influence.

By section 3.4, **MSIF** and $\mathbf{MSIF}^c$ have the same positive eigenvalues, and the eigenvector of **MSIF** can be obtained from the eigenvectors of $\mathbf{MSIF}^c$ by

$$l = \mathbf{SIF}^T l^c / \sqrt{\lambda}, \tag{5.1.8}$$

where l and l^c are respectively the eigenvectors of **MSIF** and $\mathbf{MSIF}^c$ associated with eigenvalue λ. Accordingly, we can study the largest eigenvalue of **MSIF** through the largest eigenvalue of $\mathbf{MSIF}^c$.

Theorem 5.1 *If $(y_i - \boldsymbol{x}_i\theta)\boldsymbol{x}_i^T$ for $i = 1, \ldots, n$ are independent identical normally distributed random variables with mean 0 and covariance matrix $\sigma^2 E(\boldsymbol{x}^T\boldsymbol{x})$ where $E(\boldsymbol{x}^T\boldsymbol{x})$ is a nonsingular matrix, i.e.*

$$(y_i - \boldsymbol{x}_i\theta)\boldsymbol{x}_i^T \sim N_p(0, \sigma^2 E(\boldsymbol{x}^T\boldsymbol{x})),$$

Then $\mathbf{MSIF}^c$ *has a Wishart distribution with n degrees of freedom and covariance matrix I_p, i.e. $W_p(n, I_p)$.*

Proof: Let $p \times n$ matrix Z be $(z_1, \ldots, z_n)$, where $z_i = (y_i - \boldsymbol{x}_i\theta)\boldsymbol{x}_i^T$, then

$$IF(\boldsymbol{x}_i, T, F) = \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1} z_i,$$

$$\mathbf{IF}(X, T, F) = \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1} Z,$$

and

$$\mathbf{MSIF}^c = \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1/2} Z Z^T \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1/2} / \sigma^2.$$

Since for all $i = 1, \ldots, n$,

$$z_i = (y_i - \boldsymbol{x}_i\theta)\boldsymbol{x}_i^T \sim N(0, \sigma^2 E(\boldsymbol{x}^T\boldsymbol{x})),$$

$$Z^T \sim N(0, \sigma^2 I_n \otimes E(\boldsymbol{x}^T\boldsymbol{x})).$$

Then,

$$ZZ^T \sim W_p(n, \sigma^2 E(x^T x)).$$

By properties of the Wishart distribution, we have

$$\{\sigma^2 E(x^T x)\}^{-1/2} ZZ^T \{\sigma^2 E(x^T x)\}^{-1/2} \sim W_p(n, \{\sigma^2 E(x^T x)\}^{-1/2} \sigma^2 E(x^T x) \{\sigma^2 E(x^T x)\}^{-1/2})$$

or

$$\mathbf{MSIF}^c \sim W_p(n, I_p).$$

□

Therefore, the distribution function of the largest eigenvalue λ_n of **MSIF** can be expressed in the form

$$P(\lambda_n < x) = \frac{\Gamma_p[\frac{1}{2}(p+1)]}{\Gamma_p[\frac{1}{2}(n+p+1)]} (\frac{1}{2}x)^{np/2} \cdot {}_1F_1(\frac{1}{2}n; \frac{1}{2}(n+p+1); -\frac{1}{2}xI_p), \qquad (5.1.9)$$

where

$${}_1F_1(a; c; X) = \sum_{k=0}^{\infty} \sum_{\kappa} \frac{(a)_\kappa}{(c)_\kappa} \frac{C_\kappa(X)}{k!}.$$

For a proof see p.421 of Muirhead (1982). Since the exact distribution of the largest eigenvalue λ_n is computationally difficult, it is useful to have a quick and rough approximation for its distribution function. The following bound could be used for this purpose:

$$P(\lambda_n \leq x) \leq \{P(\chi_n^2 \leq x)\}^p \qquad (5.1.10)$$

where χ_n^2 represents a random variable of chi-squared distribution with degrees of freedom n. For a proof, see p.424 of Muirhead (1982). Thus an approximate critical value for the largest eigenvalue of the **MSIF** for a least squares estimator can be set by the $(1-\alpha)^{1/p}$ quantile of a chi-squared distribution with n degrees of freedom.

In practice, the approximation given by (5.1.10) does not give strict enough bounds to decide if there are influential observations. To have an idea of the distribution of the largest

Table 5.1: 95% Quantiles of the Distribution of λ for LS estimator

		λ		λ/n	
p	n	mean	std.	mean	std.
2	25	80.56	3.05	3.222	0.122
	50	109.79	3.05	2.196	0.061
	75	143.50	3.67	1.913	0.049
	100	173.26	3.65	1.733	0.036
3	25	94.61	5.74	3.784	0.230
	50	127.40	4.50	2.548	0.090
	75	162.99	4.27	2.173	0.057
	100	194.13	4.61	1.941	0.046
4	25	113.07	7.30	4.523	0.292
	50	144.84	7.51	2.897	0.150
	75	183.41	2.91	2.445	0.039
	100	218.75	4.22	2.188	0.042

eigenvalue, we conducted a small simulation study in which we generated a random normal sample of residuals and independent variables X with $p = 2, 3, 4$ and $n = 25, 50, 75, 100$. The largest eigenvalue, λ, for each sample was calculated. Each combination was repeated for 500 times to obtain an estimate of 95% quantile. This procedure was repeated for 10 times to show the variability of the estimate. Table 5.1 lists the mean and standard deviation of the corresponding 95% quantiles of the distribution of the largest eigenvalue for the least squares estimator. For convenience of use in later examples, the quantiles divided by sample size n, λ/n, are also listed in the table.

5.2 Huber's M-estimators

Huber (1973, 1977) extended his results on robust estimation of location parameters to the case of linear regression. Huber's M-estimator T_n was defined by

$$\Gamma(T_n) = \min\{\Gamma(\theta) : \theta \in \Theta\}, \tag{5.2.1}$$

where

$$\Gamma(\theta) = \sum_{i=1}^{n} \rho((y_i - \boldsymbol{x}_i\theta)/\sigma), \tag{5.2.2}$$

for some function $\rho : \mathbf{R} \to \mathbf{R}^+$ and for a fixed σ. If ρ has derivative $(\partial/\partial r)\rho(r) = \psi(r)$, T_n satisfies the system of equations

$$\sum_{i=1}^{n} \psi((y_i - \boldsymbol{x}_i T_n)/\sigma)\boldsymbol{x}_i = 0. \tag{5.2.3}$$

Motivated by minimax asymptotic variance arguments, Huber proposed to use the function

$$\psi_c(t) = \min(c, \max(t, -c)), \tag{5.2.4}$$

where c is a constant, and usually is set to be 1.345.

In practice, we usually have to estimate the scale parameter σ along with θ. Estimators can be calculated by weighted least squares with weights (redefined iteratively) of the form

$$w_i = \min\{1, c/|r_i|\}. \tag{5.2.5}$$

It can be shown that the influence function of Huber's M-estimator for model distribution $F_\theta(\boldsymbol{x}, y)$ with density (ignoring the scale) $f_\theta(\boldsymbol{x}, y) = g(y - \boldsymbol{x}\theta)k(\boldsymbol{x})$ is given by

$$IF(\boldsymbol{x}, y; T, F_\theta) = \psi_c(y - \boldsymbol{x}\theta) \cdot M^{-1}\boldsymbol{x}^T, \tag{5.2.6}$$

where

$$\begin{aligned} M &= (E\psi_c') \cdot (E\boldsymbol{x}^T\boldsymbol{x}) \\ &= \left(\int \psi_c'(r)dG(r)\right) \cdot E(\boldsymbol{x}^T\boldsymbol{x}). \end{aligned}$$

Following Hampel (1973, 1978, 1986) we can rewrite IF as a product of two factors, namely the *influence of the residual* (IR) and the (vector-valued) *influence of position in*

factor space (IP):

$$IF(\boldsymbol{x}, r, T, F_\theta) = IR(r, T, G) \cdot IP(\boldsymbol{x}, T, K), \tag{5.2.7}$$

where

$$IR(r, T, G) = \psi_c(r)/(E\psi_c'),$$

$$IP(\boldsymbol{x}, T, K) = [E(\boldsymbol{x}^T\boldsymbol{x})]^{-1}\boldsymbol{x}^T.$$

Clearly, the influence of the residual IR is bounded. This is an improvement of the least squares estimators from a robustness point of view. However, the influence of position in factor space is unbounded. Thus an outlier in the factor space could almost completely determine the fit. In order to cope with problems caused by outlying points in the factor space, we need more refined estimators.

Similarly, the joint local influence of Huber's estimator is studied by considering the standardized influence matrix. It can be shown that

$$V(T, F) = M^{-1}QM^{-1}$$

where

$$Q = \int \psi_c^2(r)dG \cdot E(\boldsymbol{x}^T\boldsymbol{x}).$$

Thus,

$$\begin{aligned}
\mathbf{MSIF}(X, Y; T_{Huber}, F) &= IF(X, Y; T_{Huber}, F)^T V(T, F)^{-1} IF(X, Y; T_{Huber}, F) \\
&= R_\psi X Q^{-1} X^T R_\psi \qquad (5.2.8) \\
&= R_\psi X \{E(\boldsymbol{x}^T\boldsymbol{x})\}^{-1} X^T R_\psi / K_c, \qquad (5.2.9)
\end{aligned}$$

where

$$R_\psi = \text{diag}(\psi(r_1), \ldots, \psi(r_n)),$$

Table 5.2: 95% Quantiles of the Distribution of λ for Huber's estimator

		λ		λ/n	
p	n	mean	std.	mean	std.
2	25	42.94	0.79	1.718	0.032
	50	72.87	0.69	1.457	0.014
	75	103.30	1.95	1.377	0.026
	100	131.94	1.70	1.319	0.017
3	25	47.36	0.59	1.894	0.023
	50	79.96	1.06	1.599	0.021
	75	111.11	1.39	1.481	0.018
	100	141.17	1.06	1.412	0.011
4	25	50.20	0.82	2.008	0.033
	50	84.30	0.71	1.686	0.014
	75	117.62	1.24	1.568	0.017
	100	148.48	1.35	1.485	0.013

and

$$K_c = \int \psi_c^2(r) dG.$$

Since X is not bounded, the gross-error sensitivity or the single local influence is unbounded, and so is the joint local influence. However, without the normality assumption for $\psi(r)$, it would be difficult to obtain even an approximate distribution of the largest eigenvalue of **MSIF** mathematically.

As we did for the least squares estimator, we have used simulation to determine quantiles of the distribution of the largest eigenvalue for Huber's estimator. Random normal samples of residuals and independent variables X with $p = 2, 3, 4$ and $n = 25, 50, 75, 100$ were generated. The largest eigenvalue, λ, of the Huber's estimator with $c = 1.345$ for each sample was then calculated. Each combination of p and n was repeated for 500 times to obtain an estimate of the 95% quantile. This procedure was repeated for 10 times to show the variability of the estimate. Table 5.2 lists the mean and standard deviation of the 95% quantiles of the distribution of the corresponding eigenvalues. As in table 5.1, the quantiles for λ/n are listed for convenience.

5.3 Bounded Influence Estimators

Mallows (1973, 1975) proposed a way to construct bounded-influence estimators by modifying (5.2.3) to

$$\sum_{i=1}^{n} u(\boldsymbol{x}_i)\psi_c((y_i - \boldsymbol{x}_i\hat{\theta})/\sigma)\boldsymbol{x}_i^T = 0 \tag{5.3.1}$$

which, for certain weights, $u(\boldsymbol{x}_i)$ may depend on the entire X matrix and not just $\boldsymbol{x}_i$. If u is properly chosen, then the influence function will be bounded. This downweighting in X space does not include some consideration of the way the y values of these outlying observations fit in the pattern of the bulk of the data and therefore cannot be efficient.

Schweppe (Handschin, et al. 1975) proposed the form

$$\sum_{i=1}^{n} v(\boldsymbol{x}_i)\psi((y_i - \boldsymbol{x}_i\hat{\theta})/\sigma v(\boldsymbol{x}_i))\boldsymbol{x}_i^T = 0 \tag{5.3.2}$$

with $v(\boldsymbol{x}_i) = (1 - h_{ii})^{1/2}$ and $h_{ii} = \boldsymbol{x}_i(X^TX)^{-1}\boldsymbol{x}_i^T$. (5.3.2) can provide a bounded influence function and has the additional property that if $(y - \boldsymbol{x}\hat{\theta})/\sigma v(\boldsymbol{x})$ is small, the effect of $v(\boldsymbol{x})$ will be canceled out. Thus some of the efficiency problems outlined for the Mallows approach can be overcome.

The above two approaches provided bounded gross-error sensitivity. Krasker and Welsch (1982) proposed an efficient robust estimator with bounded self-standardized sensitivity. They concentrated on estimators of the form

$$\sum_{i=1}^{n} w(y_i, \boldsymbol{x}_i, T_n)(y_i - \boldsymbol{x}_i T_n)\boldsymbol{x}_i^T = 0 \tag{5.3.3}$$

where the weight function w is nonnegative, bounded, and continuous.

Let P denote a probability distribution for $(y_i, \boldsymbol{x}_i)$ on $\mathbf{R} \times \mathbf{R}^p$ satisfying $\mathcal{L}(y_i - \boldsymbol{x}_i\theta|\boldsymbol{x}_i) = N(0, \sigma^2)$ and E denote expectation with respect to P. With some regularity conditions the estimators $\{\theta_n\}$ will be consistent and asymptotically normal with influence

function

$$IF_w(y, \boldsymbol{x}, \theta, F) = w(y, \boldsymbol{x}, \theta)(y - \boldsymbol{x}\theta)B^{-1}\boldsymbol{x}^T, \tag{5.3.4}$$

where $B = -\partial/\partial\theta\{E[w(y - \boldsymbol{x}\theta)\boldsymbol{x}^T]\}_{\theta=\theta}$. An alternative expression for matrix B is

$$B = E\left[w\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right]. \tag{5.3.5}$$

It can be shown that for an estimator with weight function w, the asymptotic covariance matrix is

$$V_w = \sigma^2 E^{-1}\left[w\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right] \cdot E\left[w^2\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right] \cdot E^{-1}\left[w\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right], \tag{5.3.6}$$

and the self-standardized sensitivity is

$$\gamma_{Sw}^{**} = \sup_{y,\boldsymbol{x}}\left\{w(y, \boldsymbol{x}, \theta)^2\left(\frac{y - \boldsymbol{x}\theta}{\sigma}\right)^2 \cdot \boldsymbol{x}E^{-1}\left[w^2\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right]\boldsymbol{x}^T\right\}^{1/2}. \tag{5.3.7}$$

Now let a sensitivity bound $a > 0$ be given. Suppose there exists a weight function w that is strongly efficient among all weight function w that satisfy $\gamma_{Sw}^{**} \leq a$ and vary with y only through $|y - \boldsymbol{x}\theta|$. Then w satisfies (up to scaler multiple)

$$w(y, \boldsymbol{x}, \theta, \sigma, A) = \min\left\{1, \frac{a}{\left|\frac{y-\boldsymbol{x}\theta}{\sigma}\right|\left\{\boldsymbol{x}E^{-1}\left[w^2\left(\frac{\epsilon}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right]\boldsymbol{x}^T\right\}^{1/2}}\right\} \tag{5.3.8}$$

or

$$w(y, \boldsymbol{x}, \theta, \sigma, A) = \min\left\{1, \frac{a}{\left|\frac{y-\boldsymbol{x}\theta}{\sigma}\right|\{\boldsymbol{x}A^{-1}\boldsymbol{x}^T\}^{1/2}}\right\}, \tag{5.3.9}$$

where

$$A = E\left[w(y, \boldsymbol{x}, \theta, \sigma, A)^2\left(\frac{y - \boldsymbol{x}\theta}{\sigma}\right)^2 \boldsymbol{x}^T\boldsymbol{x}\right]. \tag{5.3.10}$$

This leads to the Krasker-Welsch estimators $(\theta_n, \sigma_n, A_n)$ which satisfy

$$\frac{1}{n}\sum_{i=1}^n \min\left\{1, \frac{a}{\left|\frac{y_i - \boldsymbol{x}_i\theta_n}{\sigma_n}\right|\{\boldsymbol{x}_i A_n^{-1}\boldsymbol{x}_i^T\}^{1/2}}\right\} \cdot (y_i - \boldsymbol{x}_i\theta_n)\boldsymbol{x}_i^T = 0 \tag{5.3.11}$$

$$A_n = \frac{1}{n}\sum_{i=1}^n g_1\left(\frac{a}{(\boldsymbol{x}_i A_n^{-1}\boldsymbol{x}_i^T)^{1/2}}\right)\boldsymbol{x}_i^T\boldsymbol{x}_i, \tag{5.3.12}$$

where

$$g_1(t) = E_z \min\{z^2, t^2\}$$

and z is an $N(0,1)$ random variable independent of $\boldsymbol{x}$.

Krasker and Welsch described an approach to computing Krasker-Welsch estimates and discussed the problem of choosing a, the bound on the self-standardized sensitivity. The important feature of the estimator is that it satisfies a first-order condition for strong efficiency subject to the constraint.

As for Huber's estimator, the standardized influence matrix for the KW estimator, T_{KW} is given by

$$\mathbf{MSIF}(X, Y; T_{KW}, F) = W_a R X A^{-1} X^T R W_a / \sigma^2 \tag{5.3.13}$$

where $W_a = \text{diag}(w_1, \ldots, w_n)$, $R = \text{diag}(r_1, \ldots, r_n)$ and $r_i = y_i - \boldsymbol{x}_i\theta, i = 1, \ldots, n$. Since the Krasker and Welsch estimator has bounded self-standardized sensitivity (single local influence), for a given n, it has bounded local influence too.

As before, it is almost impossible to obtain the distribution of the largest eigenvalue of **MSIF** for the KW estimator mathematically. Hence we have conducted a small simulation study to determine the quantiles of the distribution of the largest eigenvalue for the KW estimator with $a = 3.2$. Random normal samples of residuals and independent variables X with $p = 2, 3, 4$ and $n = 25, 50, 75, 100$ were generated. The largest eigenvalue, λ, of **MSIF** for the KW estimator with $a = 3.2$ for each sample was then calculated. Each combination of p and n was repeated for 500 times. This procedure was repeated for 10 times to show the variability of the estimate. Table 5.3 lists the mean and standard deviation of the 95% quantiles of the distribution of the corresponding eigenvalues for the KW estimator. As in table 5.1-5.2, the quantiles for λ/n are also listed for convenience.

Table 5.3: 95% Quantiles of the Distribution of λ for KW estimator

		λ		λ/n	
p	n	mean	std.	mean	std.
2	25	38.82	0.62	1.553	0.025
	50	67.70	0.98	1.354	0.020
	75	96.07	0.78	1.281	0.010
	100	124.16	0.84	1.242	0.008
3	25	42.82	0.57	1.713	0.023
	50	72.93	0.68	1.459	0.014
	75	103.18	0.78	1.376	0.010
	100	131.33	1.14	1.313	0.011
4	25	46.25	0.41	1.850	0.017
	50	77.91	1.03	1.558	0.021
	75	108.90	1.28	1.452	0.017
	100	138.55	1.00	1.386	0.010

5.4 High Breakdown-point Estimators

The degree of robustness of an estimate in the presence of outliers can also be measured by the breakdown point. Many of the proposed robust estimators in regression fail to have high breakdown point. In recent years, several estimators with high breakdown point, i.e., 0.5, were proposed.

Recall that in the formation of the standardized influence matrix, it is essential to replace the true parameter with a global robust estimator to study the joint local influence of one particular dataset. In a linear regression framework, we recommend using the high breakdown point estimator that is available.

5.4.1 The Least Median Squares (LMS) Estimator

Rousseeuw (1984) proposed the *least median squares* (LMS) estimator which is defined by the minimization of the median of the squares of the residuals. This yields the estimator $\hat{\theta}$ given by

$$\min_{\hat{\theta}} \text{med}_i r_i^2 \tag{5.4.1}$$

where $r_i = y_i - x_i\hat{\theta}$ are the residuals. It turns out that this estimator is very robust with respect to outliers in y as well as in x. It was shown that its breakdown point is 50%. The algorithms for computation of LMS estimates are discussed in Leroy and Rousseeuw (1984) and Steele and Steiger (1984).

The scale estimate of LMS estimator is calculated by the following. First define

$$s_0 = 1.4826\left(1 + \frac{5}{n-p}\right)\sqrt{\text{med}_i r_i^2}, \tag{5.4.2}$$

then a weight w_i for ith observation is determined by

$$w_i = \begin{cases} 1 & \text{if } |r_i/s_0| \leq 2.5 \\ 0 & \text{otherwise} \end{cases} . \tag{5.4.3}$$

Finally,

$$\hat{\sigma}^{*2} = \frac{\sum_{i=1}^n w_i r_i^2}{\sum_{i=1}^n w_i - p}. \tag{5.4.4}$$

Note that $\hat{\sigma}^*$ also has a 50% breakdown point.

A disadvantage of the LMS estimator is its lack of efficiency when the errors are really normally distributed. And Rousseeuw and Leroy (1987) have shown that it has an abnormally slow convergence rate. To repair this, Rousseeuw (1983, 1984) introduced the *least trimmed squares* (LTS) estimator, given by

$$\min_{\hat{\theta}} \sum_{i=1}^{h} (r^2)_{(i)} \tag{5.4.5}$$

where $(r^2)_{(1)} \leq \ldots \leq (r^2)_{(n)}$ are the ordered squared residuals. Generalizing this, Rousseeuw and Yohai (1984) introduced the *S-estimator*, corresponding to

$$\min_{\hat{\theta}} S(\theta) \tag{5.4.6}$$

where $S(\theta)$ is a certain type of robust M-estimate of the residuals $r_1, \ldots, r_n$. Like LMS estimators, both LTS estimators and S-estimators have a 50% breakdown point by a suitable

choice of the constants involved. But they converge at the usual rate. On the other hand, the efficiency problem can also be overcome by calculating a one-step M-estimator using the LMS estimator as the starting point.

Besides the efficiency problem of the LMS estimator, Hettmansperger and Sheatter (1992) pointed out that the LMS estimator is sensitive to inliers, and Davis (1993) showed that it has bounded influence only under certain conditions. Nevertheless, as a starting estimator for constructing the standardized influence matrix, the LMS estimator or LTS estimator is recommended as long as consistency and a global robust property are the main concerns.

5.4.2 Median Absolute Deviation (MAD)

Even though the estimator of σ^2 is not the primary interest of the problem, a robust estimator is needed to quantify the standardized influence matrix.

The median absolute deviation (MAD) estimator is one of commonly used estimator which has maximal breakdown point $\epsilon^* = 0.5$. It is defined as

$$\hat{\sigma} = 1.483\text{MAD}(r_i) = 1.483\text{med}_i\{|r_i - \text{med}_j(r_j)|\}. \tag{5.4.7}$$

5.4.3 Robust Estimator for $E(\boldsymbol{x}^T\boldsymbol{x})$

One way of obtaining a robust estimate for $E(\boldsymbol{x}^T\boldsymbol{x})$ is to use the weights given by LMS estimator. It can be calculated by the following,

$$\hat{E}(\boldsymbol{x}^T\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \boldsymbol{x}_i^T \boldsymbol{x}_i, \tag{5.4.8}$$

where ω_i is given by (5.4.3).

There is another robust estimator that is called minimum volume ellipsoid (MVE) estimator. Rousseeuw (1983, 1984) introduced an estimator with maximal breakdown point

given by

$$\mu(\boldsymbol{x}) = \text{ center of the minimal volume ellipsoid}$$
$$\text{covering (at least) h points of x,} \tag{5.4.9}$$

where h can be taken equal to $[n/2]+1$. The corresponding covariance estimator $C(\boldsymbol{x})$ is given by the ellipsoid itself, multiplied by a suitable factor to obtain consistency. Rousseeuw (1986) also showed that the breakdown point of a MVE estimator converges to 50% as $n \to \infty$. An algorithm to obtain the MVE estimator can be found in Rousseeuw (1986).

Therefore,

- in the case of where no intercept is included, a global robust estimator of $E(\boldsymbol{x}^T\boldsymbol{x})$ could be determined by

$$\hat{E}(\boldsymbol{x}^T\boldsymbol{x}) = C(\boldsymbol{x}) \tag{5.4.10}$$

  where $C(\boldsymbol{x})$ is determined by MVE estimator.

- with an intercept, a global robust estimator of $E(\boldsymbol{x}^T\boldsymbol{x})$, where $\boldsymbol{x} = (1, \tilde{\boldsymbol{x}})$, could be determined by

$$\hat{E}(\boldsymbol{x}^T\boldsymbol{x}) = \begin{pmatrix} 1 & 0 \\ 0 & C(\tilde{\boldsymbol{x}}) \end{pmatrix} \tag{5.4.11}$$

  where $C(\tilde{\boldsymbol{x}})$ is determined by the MVE estimator applying to $\tilde{\boldsymbol{x}}$.

Since MVE procedure assumes the ellipsoid distribution for $\boldsymbol{x}$, it is limited useful in estimating $\hat{E}(\boldsymbol{x}^T\boldsymbol{x})$.

Chapter 6

Logistic Regression Applications

In this chapter the standardized influence matrix is applied in a logistic regression framework. First, the standardized influence matrix is developed for a maximum likelihood estimator. Then the same procedure has been applied to Beran's robust estimator, Morgenthaler's least absolute deviation estimator and Stefanski's bounded influence estimator. Finally, a global robust estimator for grouped data is proposed to construct a local influence measure for estimators. Special action is needed in dealing with logistic regression when data are ungrouped. This will be discussed in the next chapter.

We consider the logistic model defined in chapter 2. Let y_i given $\boldsymbol{x}_i$ follow a binomial distribution with parameter m and probability $\pi(\boldsymbol{x}_i)$ such that

$$\pi(\boldsymbol{x}_i) = \tau(\boldsymbol{x}_i\theta),\ i = 1, \ldots, n, \tag{6.0.1}$$

where $\boldsymbol{x}_i \in \mathbf{R}^p$ is the ith row of the design matrix $X_{n\times p}$, $\theta \in \Theta \subset \mathbf{R}^p$ is a p-vector of unknown parameters ($p \geq 1$) and

$$\tau(t) = (1 + \exp(-t))^{-1}. \tag{6.0.2}$$

Also, we assume the joint distribution F of $(\boldsymbol{x}, y)$ has a density of the form

$$f(\boldsymbol{x}, y; \theta) = g(y; \boldsymbol{x}\theta)k(\boldsymbol{x}) \tag{6.0.3}$$

6.1 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator of θ is obtained by solving the equation

$$\sum_{i=1}^{n}(y_i/m - \tau(\boldsymbol{x}_i\theta))\boldsymbol{x}_i^T = 0 \tag{6.1.1}$$

and is, under certain conditions (see Agresti 1990, chapter 12), optimal in the sense of asymptotically efficient. The functional $T(F)$ corresponding to the MLE is the solution to

$$\int (y/m - \tau(\boldsymbol{x} \cdot T(F)))\boldsymbol{x}^T dF(\boldsymbol{x}, y) = 0 \tag{6.1.2}$$

where $F(\boldsymbol{x}, y)$ is the joint distribution of $(\boldsymbol{x}, y)$. Let $M = E(w(\boldsymbol{x}\theta)\boldsymbol{x}^T\boldsymbol{x})$, where $w(t) = \tau(t)(1 - \tau(t))$. The influence function of $T(F)$ is

$$IF(\boldsymbol{x}, y; T, F) = M^{-1}(y/m - \tau(\boldsymbol{x}\theta))\boldsymbol{x}^T. \tag{6.1.3}$$

Since

$$V(T, F) = \frac{1}{m}M^{-1}, \tag{6.1.4}$$

$$SIF(\boldsymbol{x}, y; T, F) = \frac{1}{\sqrt{m}}M^{-1/2}(y/m - \tau(\boldsymbol{x}\theta))\boldsymbol{x}^T \tag{6.1.5}$$

and

$$\mathbf{MSIF}(X, Y; T, F) = mRXM^{-1}X^TR, \tag{6.1.6}$$

where $R = diag(r_1, \ldots, r_n)$ and $r_i = y_i/m - \tau(\boldsymbol{x}_i\theta)$ for $i = 1, \ldots, n$.

Although r_i is bounded, the influence function or gross-error sensitivity is unbounded as $\boldsymbol{x}$ is unbounded. Therefore, the MLE has unbounded single local influence as well as unbounded joint local influence. In other words, the MLE in logistic regression can be highly influenced by a relatively small subset of the data.

Unlike the linear regression case where the distribution of the residuals r_i is free of parameters, the distribution of $r_i = y_i/m - \tau(\boldsymbol{x}_i\theta)$ in logistic regression depends on the unknown parameters. Consequently, the distribution function of the largest eigenvalue of the standardized influence matrix is a function of the parameter θ, and the exact critical value of the largest eigenvalue is also a function of θ. However, the following approximation approach may apply to some situations:

- For all m, the np-guideline proposed in chapter 4 can always serve as a reasonable reference.
- When m is large enough, say $m \geq 10$, the m-asymptotic distribution of the standardized influence matrix is a Wishart distribution. Thus, the discussion in section 5.1 applies for this case.
- Even when all above fails, one can, based on a global robust estimator of the parameter, do a small simulation study to obtain the critical value of the largest eigenvalue for the specific regression coefficient.

6.2 Beran's Robust Estimator

Beran (1982) proposed a general theory for robust estimation in independent non-identical distributed (i.n.i.d) models, which avoids considerations of efficiency. Following is a summary of his theory only as it pertains to logistic regression. Start by assuming that the true distribution of y_i given $\boldsymbol{x}_i$ satisfies

$$E(y_i/m|\boldsymbol{x}_i) = P_i, \ i = 1, \ldots, n. \tag{6.2.1}$$

Define $\hat{\theta}$ as the value of θ which minimizes

$$\sum_{i=1}^{n} \|P_i - \tau(\boldsymbol{x}_i\theta)\|^2_{i,\theta}, \tag{6.2.2}$$

where the choice of the norm $\|\cdot\|_{i,\theta}$ is made by the statistician. Beran suggests taking $\|\cdot\|_{i,\theta} = |\cdot|$ and for this choice $\hat{\theta}$ exists uniquely. The value $\hat{\theta}$ minimizes the distance between the true distribution of the response $(P_1, \ldots, P_n)$ and the modeled distribution $(\tau(x_1\theta), \ldots, \tau(x_n\theta))$.

Since Beran's theory distinguishes estimators only via their asymptotic properties, Stefanski (1983) concluded that the value of θ which minimizes

$$\sum_{i=1}^{n}(y_i/m - \tau(x_i\theta))^2 \tag{6.2.3}$$

performs optimally in Beran's framework. Let $w(x_i\theta) = \tau'(x_i\theta)$. The resulting estimator is a solution to

$$\sum_{i=1}^{n}(y_i/m - \tau(x_i\theta))w(x_i\theta)x_i^T = 0. \tag{6.2.4}$$

The contamination neighborhood over which this estimator is optimal takes the form

$$B_n(\theta_0, C) = \{P_1, \ldots, P_n : \sum_{i=1}^{n}(P_i - \tau(x_i\theta))^2 \leq 2C^2\}. \tag{6.2.5}$$

The solution to (6.2.4) is none other than the least squares estimator of θ. In fact, for binary (binomial) regression, minimizing the distance between the observed or empirical distribution and the modeled distribution is equivalent to minimizing the distance between the observed regression function and the model regression function.

The functional $T(F)$ corresponding to the estimator determined by (6.2.4) is the solution to

$$\int (y/m - \tau(x \cdot T(F)))w(x \cdot T(F))x^T dF(x, y) = 0.$$

Let

$$M = E[(w(x \cdot T(F)))^2 x^T x]$$

and

$$Q = E[(w(x \cdot T(F)))^3 x^T x].$$

The influence function is

$$IF(\boldsymbol{x}, y; T, F) = M^{-1}(y - \tau(\boldsymbol{x} \cdot T(F))w(\boldsymbol{x} \cdot T(F))\boldsymbol{x}^T, \tag{6.2.6}$$

and

$$V(T, F) = M^{-1}QM^{-1}/m. \tag{6.2.7}$$

The standardized influence matrix can be written as

$$\mathbf{MSIF}(X, Y; T, F) = mRWXQ^{-1}X^TWR \tag{6.2.8}$$

where $R = \text{diag}(r_1, \ldots, r_n)$, $r_i = y_i/m - \tau(\boldsymbol{x}_i \cdot T(F))$ for $i = 1, \ldots, n$; and $W = \text{diag}(w(\boldsymbol{x}_1 \cdot T(F)), \ldots, w(\boldsymbol{x}_n \cdot T(F)))$.

Since $w(\boldsymbol{x} \cdot T(F))\boldsymbol{x}^T \to 0$ as $|\boldsymbol{x}| \to \infty$ when $T(F) \neq 0$, the factor $w(\boldsymbol{x}\theta)$ limits the influence over much of the design space. And the influence function of Beran's estimator is bounded for all $\boldsymbol{x}$ in logistic regression with $T(F) \neq 0$ except at the intercept. Therefore, generally speaking, the gross-error sensitivity of Beran's estimator is bounded, so is the joint local influence for a given n.

6.3 Least Absolute Deviation Estimator

Morgenthaler (1992) proposed a least absolute deviation estimator by replacing the L_2-norm by the L_1-norm in the deviation of quasi-likelihoods.

A quasi-likelihood function of a generalized linear model is defined by its gradient,

$$\frac{\partial}{\partial \mu} K(\mu; Y) = W(\mu)^{-1}(Y - \mu), \tag{6.3.1}$$

where $\mu = (\mu_1, \ldots, \mu_n)^T$ is the mean of Y, and $W(\mu)$ denotes the variance matrix of the vector of responses. Then the quasi-likelihood estimator is the solution to the equations

$$U(\theta) = D^TW(\mu)^{-1}(Y - \mu) = 0. \tag{6.3.2}$$

The matrix D contains the partial derivatives $\partial\mu_i/\partial\theta_j = d_{ij}$.

With a constant diagonal scatter matrix, $W = \text{diag}(w_1, \ldots, w_n)$, fitting by least L_q-norm ($q \geq 1$) corresponds to minimizing

$$\sum_{i=1}^{n} \left| \frac{y_i - \mu_i}{w_i^{1/2}} \right|^q .$$

The corresponding gradient is

$$\{\text{diag}(w_1, \ldots, w_n)\}^{-q/2}\{|Y - \mu|^{q-1}\text{sgn}(Y - \mu)\}, \tag{6.3.3}$$

where sgn(.) denotes the element-wise sign-function. With $q = 2$, this formula reduces to the quasi-likelihood in (6.3.1). However, this extension does not lead to consistent estimates unless $q = 2$ or unless the responses are symmetrically distributed around their means. One must correct for asymmetry and that demands detailed knowledge of the underlying distribution.

To obtain consistent estimates Morgenthaler calculated the correcting quantities

$$c_i = E\{|y_i - \mu_i|^{q-1}\text{sgn}(y_i - \mu_i)\}. \tag{6.3.4}$$

The corrected gradient is

$$\{\text{diag}(w_1, \ldots, w_n)\}^{-q/2}\{|Y - \mu|^{q-1}\text{sgn}(Y - \mu) - c\}, \tag{6.3.5}$$

and the estimating equation for θ is

$$U_q(\theta) = D^T\{\text{diag}(w_1, \ldots, w_n)\}^{-q/2}\{|Y - \mu|^{q-1}\text{sgn}(Y - \mu) - c\}. \tag{6.3.6}$$

In binary logistic regression, the correction for consistency is

$$c_i = (1 - \mu_i)^{q-1}\mu_i - \mu_i^{q-1}(1 - \mu_i).$$

This leads to the estimating equation,

$$U_q(\theta) = D^T W^{-q/2}[(Y - \mu)\{(1 - \mu)^{q-1} + \mu^{q-1}\}]. \tag{6.3.7}$$

Since $D = WX$, the least absolute deviation estimator for this particular model is the solution to

$$X^T W^{1/2}(Y - \mu) = \sum_{i=1}^{n} \{\mu_i(1 - \mu_i)\}^{1/2}(y_i - \mu_i)x_i^T = 0. \tag{6.3.8}$$

It can be shown that the influence function of the least absolute deviation estimator for a binary logistic model is

$$IF(x, y; T, F) = M^{-1}\{w(x \cdot T(F))\}^{1/2}(y - \tau(x \cdot T(F)))x^T \tag{6.3.9}$$

where

$$M = E(\{w(x \cdot T(F))\}^{3/2}x^T x)$$

and $\mu = \tau(x \cdot T(F))$. Now let

$$Q = E(\{w(x \cdot T(F))\}^2 x^T x)$$

and we have

$$V(T, F) = M^{-1}QM^{-1}. \tag{6.3.10}$$

Thus, the standardized influence matrix can be written as

$$\mathbf{MSIF}(X, Y; T, F) = RW^{1/2}XQ^{-1}X^T W^{1/2}R. \tag{6.3.11}$$

Like Beran's robust estimator, the weight $\{w(x\theta)\}^{1/2}$ limits the influence of the design space. Therefore, the least absolute deviation estimator has bounded gross-error sensitivity or single local influence. Consequently, the joint local influence for a given n is also bounded.

6.4 Bounded Influence Estimator

Stefanski, et al (1986) extended the Krasker-Welsch estimator for a linear model and proposed an optimal bounded influence estimator for generalized linear models.

They start with considering M-estimator $\hat{\theta}_\psi$ satisfying

$$\sum_{i=1}^{n} \psi(x_i, y_i, \hat{\theta}_\psi) = 0, \tag{6.4.1}$$

with

$$E_\theta(\psi(x, y, \theta)) = 0. \tag{6.4.2}$$

Under regularity conditions, $\hat{\theta}_\psi$ is consistent and asymptotically normal with influence function

$$IF(x, y; \theta(F)) = D_\psi^{-1}(\theta)\psi(x, y, \theta), \tag{6.4.3}$$

where

$$D_\psi(\theta) = -\frac{\partial}{\partial\beta}[E_\theta(\psi(x, y, \beta))]_{\beta=\theta}. \tag{6.4.4}$$

Now let

$$Q_\psi(\theta) = E_\theta(\psi(\theta)\psi^T(\theta))$$

and θ_0 be the true unknown parameter. The asymptotic variance of $n^{1/2}(\hat{\theta}_\psi - \theta_0)$ is

$$V(T, F) = D_\psi^{-1}(\theta_0)Q_\psi(\theta_0)D_\psi^{-1}(\theta_0)^T. \tag{6.4.5}$$

Bounded influence estimators are those M-estimators $\hat{\theta}_\psi$ with $s(\psi) \leq a < \infty$, where $s(\psi)$ is the self-standardized sensitivity of the estimator $\hat{\theta}_\psi$. The score function of the optimal bounded influence estimator is

$$\psi_{BI}(x, y, \theta) = (l - c)\min^{1/2}[1, a^2/\{(l - c)^T B^{-1}(l - c)\}], \tag{6.4.6}$$

where $l = l(x, y, \theta)$ is the score function of the maximum likelihood estimator ($(y/m - \tau(x\theta))x^T$ for logistic regression), and the $p \times 1$ vector $c = c(\theta)$ and the $p \times p$ matrix $B = B(\theta)$ are functions of θ defined implicitly by the equations

$$E_\theta[\psi_{BI}(x, y, \theta)] = 0,\ B(\theta) = E_\theta(\psi_{BI}\psi_{BI}^T). \tag{6.4.7}$$

With $c(\theta)$ and $B(\theta)$ so defined, ψ_{BI} is unbiased and $Q_{\psi_{BI}}(\theta) = B(\theta)$. Thus, θ_{BI} has bounded self-standardized sensitivity with bound a.

Stefanski also showed that for a given bound $a > 0$, if there exists an unbiased, strongly efficient score ψ_{opt} satisfying $s(\psi) \leq a < \infty$, then ψ_{opt} is equivalent to ψ_{BI} whenever the latter is defined. This result suggests the estimator $\hat{\theta}_{BI}$ be obtained by solving

$$\sum_{i=1}^{n} \hat{\psi}_i(\hat{\theta}_{BI}) = 0, \tag{6.4.8}$$

where $\hat{\psi}_i(\theta) = \psi_{BI}(\boldsymbol{x}_i, y_i, \theta, \hat{B}(\theta), \hat{c}(\theta))$ and $\hat{c}(\theta)$ and $\hat{B}(\theta)$ are defined implicitly by the equations

$$\sum_{i=1}^{n} E_\theta(\hat{\psi}_i(\theta)) = 0, \;\; \hat{B}(\theta) = n^{-1} \sum_{i=1}^{n} E_\theta(\hat{\psi}_i(\theta)\hat{\psi}_i^T(\theta)). \tag{6.4.9}$$

Since solving (6.4.8) and (6.4.9) for logistic regression is difficult, Stefanski suggested substituting a one step approximation with a bounded leverage estimator as the initial estimator. See Stefanski (1986) for the details of the estimation procedure.

With known $c(\theta)$ and $B(\theta)$, the influence function of the bounded influence estimator is

$$IF(\boldsymbol{x}, y; \theta_{BI}, F) = D_{\psi_{BI}}^{-1}(\theta) w(\boldsymbol{x}, y, \theta)(l(\boldsymbol{x}, y, \theta) - c(\theta)), \tag{6.4.10}$$

where $w(\boldsymbol{x}, y, \theta) = \min^{1/2}[1, a^2/\{(l(\boldsymbol{x}, y, \theta) - c(\theta))^T B^{-1}(l(\boldsymbol{x}, y, \theta) - c(\theta))\}]$. Moreover the standardized influence matrix can be written as

$$\mathbf{MSIF}(X, Y; \theta_{BI}, F) = W S_X B^{-1} S_X W, \tag{6.4.11}$$

where the $n \times p$ matrix

$$S_X = \text{diag}(l(\boldsymbol{x}_1, y_1, \theta) - c(\theta), \ldots, l(\boldsymbol{x}_n, y_n, \theta) - c(\theta))$$

and $W = \text{diag}(w(\boldsymbol{x}_1, y_1, \theta), \ldots, w(\boldsymbol{x}_n, y_n, \theta))$.

Obviously, the estimator has bounded gross-error sensitivity or bounded single local influence. For a given n, it has bounded joint local influence.

6.5 Globally Robust Estimator for Grouped Data

As in the case of linear regression, it is important to replace the true parameter with a global robust estimator to study joint local influence. The following proposed robust estimator for grouped data is an extension of the least median squares estimator for the linear model.

6.5.1 Empirical Logistic Transformation

Suppose that Y follows a binomial distribution with parameter m and probability p, and suppose that we require an approximately unbiased estimate of the log odds,

$$z(p) = \log\left(\frac{p}{1-p}\right). \tag{6.5.1}$$

Provided that neither the number of successes nor the number of failures is too small, it is natural that $z(p)$ be estimated by

$$Z = \log(\left(\frac{Y + c_0}{m - Y + c_0}\right) \tag{6.5.2}$$

where c_0 is a constant. The maximum likelihood estimator has this form with $c_0 = 0$ and has asymptotic bias of order $O(m^{-1})$. For the particular choice $c_0 = 1/2$, the estimator $Z_{1/2}$ has asymptotic bias of order $O(m^{-2})$. This is known as the empirical logistic transformation (Cox 1970). Cox (1989) and Gart (1967, 1985, 1986) discussed several issues about the variance estimator of Z and applications of empirical logistic regression.

With the empirical logistic transformation we expect $E(Z) = X\theta$ and we can use linear regression techniques to obtain the parameter estimator. However, as Gart (1986)

pointed out,

- $E(Z) \neq z(p)$ in finite samples whenever $p \neq 1/2$, even for the optimal value of $c = 1/2$.
- An unbiased estimate of $\mathrm{Var}(Z)$ cannot be found in a finite sample.
- The distribution of Z deviates appreciatively from normality even for a fairly large sample size.
- Z and its estimated variance are highly correlated.

Thus, the transformation is useful only if all the binomial indices m are fairly large.

6.5.2 Empirical Logistic Least Median Squares Estimator

Despite the disadvantages of using the empirical logistic transformation, the least median squares estimator of the transformed linear model where

$$E(Z) = z(p) = X\theta \tag{6.5.3}$$

is a good starting point and is a highly robust estimator for the logistic model. However, a straight application of LMS results in biased estimates because of nonlinearity between x and the transformed values of the extreme responses (0 or m). The following proposed two stage robust estimator is based on the assumption that there is enough non-extreme responses by which θ is mostly determined.

Stage I: Empirical logistic least median squares.

First of all, the responses are transformed by (6.5.2) with $c = 1/2$. Based on the assumption made, a rough estimate $\hat{\theta}^0$ of θ is determined by LMS for model (6.5.3) for non-extreme responses only. Then the weight (inverse of the variance) w^0 for each observation can be calculated with $\hat{\theta}^0$. Finally, $\hat{\theta}^1$ is obtained by LMS for model (6.5.3) incorporated with weight w^0.

1. Apply the empirical logistic transformation to the response to generate Z, where

$$z_i = \begin{cases} -\log(2m+1) & \text{if } y_i = 0; \\ \log\left(\frac{y_i+1/2}{m-y_i+1/2}\right) & \text{if } y_i = 1, \ldots, m-1; \\ \log(2m+1) & \text{if } y_i = m. \end{cases}$$

2. The LMS procedure is applied to the model

$$E(Z^0) = X^0\theta,$$

where (X^0, Z^0) are the transformed observations with non-extreme responses. Specifically, $\hat{\theta}^0$ is the solution to

$$\min_{\theta} \text{med}_{1 \le y_i \le m-1}(z_i - \boldsymbol{x}_i\theta)^2. \tag{6.5.4}$$

3. The weight w^0 is calculated by

$$w_i^0 = \frac{1}{p_i(\hat{\theta}^0) \cdot (1 - p_i(\hat{\theta}^0))}.$$

4. The LMS procedure is applied to the model

$$E(Z^1) = X^1\theta,$$

where (X^1, Z^1) are the transformed observations whose weights are larger than a small constant, b; that is, $\hat{\theta}^1$ is determined by

$$\min_{\theta} \text{med}_{w_i^0 > b}(z_i - \boldsymbol{x}_i\theta)^2. \tag{6.5.5}$$

The choice of the constant b is a tradeoff between the bias and stability of the estimates. For smaller b, the estimates are more biased but more stable. Similarly, we can use a LTS procedure instead of a LMS procedure to obtain more stable estimates.

Stage II: One step improvement

Because of the disadvantages of the empirical logistic transformation, the first stage estimator needs improvements. In linear regression, the one step M-estimator with LMS not only keeps the high breakdown property of the LMS but also has good efficiency. Likewise, a certain form of the one step M-estimator used in the first stage estimator should gain a good deal of efficiency and maintain the global robustness property.

Similar to the algorithm for the Least Absolute Deviations estimator, the one step estimator is obtained by

$$\theta_1 = \theta_0 + (X^T M X)^{-1} X^T W (y/m - \mu) \tag{6.5.6}$$

where M is diagonal with elements

$$m_{ii} = w(\boldsymbol{x}_i\theta_0)\{w(\mu_i) - w'(\mu_i)(y_i/m - \mu_i)\},$$

and $w(\mu_i) = \{\mu_i(1-\mu_i)\}^{1/2}$.

One possible extension to the above one step estimator is to make $r = y/m - \mu$ be bounded further so that possible outliers would not have too much influence on the one step improvement. We recommend using the Huber function that is defined by

$$\psi_c(r) = \begin{cases} r, & \text{if } \frac{|r|}{\mu(1-\mu)/m} \leq c; \\ \text{sgn}(r)c, & \text{otherwise,} \end{cases} \tag{6.5.7}$$

with $c = 1.345$ as suggested by Huber.

Therefore, the one step improvement proposed is defined by

$$\hat{\theta}^2 = \hat{\theta}^1 + (X^T M_c X)^{-1} X^T W \psi_c(r) \tag{6.5.8}$$

where M_c is diagonal with elements

$$m_{ii} = w(\boldsymbol{x}_i\hat{\theta}^1)\{w(\hat{\mu}_i) - w'(\hat{\mu}_i)\psi_c(r_i)\},\ i = 1, \ldots, n.$$

6.5.3 High Breakdown Property and Consistency

In the context of logistic regression, the concept of breakdown needs to be revised. It is obvious that the parameter estimator may not take arbitrary large values. It makes more sense to examine probability estimates than parameter estimates. Therefore, a reasonable definition of breakdown in the logistic regression context ought to be when all the probability estimates tends to the extreme values 0 or 1.

Since the LMS estimator in linear regression has breakdown point 0.5, for a given l, $\hat{\theta}^1$ won't breakdown unless the response of more than $n/2 - l$ data points changes to the extreme value (0 or m). Thus it has breakdown point $0.5 - l/n$. Consequently, $\hat{\theta}^2$ also has breakdown point $0.5 - l/n$. Accordingly, we conclude that the breakdown point of $\hat{\theta}^1$ or $\hat{\theta}^2$,

$$\epsilon^* = \frac{1}{2} - \frac{l}{n}, \tag{6.5.9}$$

where l is maximum number of extreme responses, which depends on $X\theta$.

Generally speaking, with regular non-extreme X and θ, we expect that l is around $n/(m+1)$. Therefore, the breakdown point of the two stage estimator is

$$\epsilon^* \approx \frac{1}{2} - \frac{1}{m+1}. \tag{6.5.10}$$

Note that ϵ^* tends to be 0.5 if m is large enough; and $\epsilon^* \approx 0$ in the case of ungrouped data where $m = 1$, which suggests that special attention need to be paid in the case of ungrouped data.

In terms of consistency, the two stage estimator, $\hat{\theta}^2 \to \theta_0$ in probability as $m, n \to \infty$. However, for moderate $5 \leq m \leq 10$, the estimator Z may not be even approximately consistent because of the weakness of the empirical logistic transformation. This results in the

inconsistency of $\hat{\theta}^1$ for moderate m, which in fact is the reason for a one step improvement. The one step improvement makes the estimates approximately consistent for many cases, as will be shown by the simulation studies described in a later chapter.

Chapter 7

Binary Logistic Regression

As it has been shown in previous chapters, the work done in linear regression provides a guide or methodology that may be applied to the logistic regression model. However, as Jennings (1986) pointed out, the similarities that allow adaptation of such techniques to logistic regression seem to hide many of the differences. This chapter discusses some of the differences and difficulties specifically in the case of binary data with no replicates.

7.1 Residuals and Outliers

There are several ways of defining residuals in logistic regression as shown in chapter 2. Many other definitions of generalized residuals for the generalized linear model have been discussed by Pierce and Schafer (1986). Based on m-asymptotic, they used generalized residuals to identify individual poorly fitting observations. However, in the case of binary logistic regression where $m = 1$, the asymptotic results no longer hold.

Obviously, the major difficulty in binary logistic regression is that the deviance or Pearson residuals have a two point distribution that depends on p. For instance, the deviance residual is

$$r_{Di} = \{-2(y_i \log p_i + (1 - y_i) \log(1 - p_i))\}^{1/2},$$

and the distribution of r_{Di} depends on p_i. Specifically, $Er_{Di} \neq 0$ and Er_{Di}^2 are not constant

– both depend on p_i (see Jennings 1986). Consequently, the distribution function of the residuals is a function of regression coefficients θ. Therefore, the usual concept of outliers based on residuals or the slippage model no longer works.

Jennings (1986) proposed a reasonable definition of an outlier by looking at the probability of the response Y being very far away from the true mean EY. Formally, the ith case is an outlier if

$$\Pr(|Y_i - EY_i| \geq |y_i - EY_i|) < \epsilon,$$

where ϵ is small.

According to the definition above, outliers in the binary regression context can only happen where EY_i or $1 - EY_i$ is close enough to 0 or 1. Now suppose a contaminated sample consists of $n-k$ observations $Z = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{n-k}, y_{n-k})\}$ from postulated model $EY = \tau(\boldsymbol{x}\theta)$ with unknown θ and k other observations $Z_c = \{(\boldsymbol{x}_{n-k+1}, y_{n-k+1}), \ldots, (\boldsymbol{x}_n, y_n)\}$. Then

- $(\boldsymbol{x}_l, y_l), l = n - k + 1, \ldots, n$ is indistinguishable from Z, the correct samples, if $\tau(\boldsymbol{x}_l\theta)$ is not close to the boundary of p, i.e. 0 or 1.
- The observation in Z_c could be identified as an outlier only if y is close to $1 - \tau(\boldsymbol{x}\theta)$ with boundary value of $\tau(\boldsymbol{x}\theta)$.

7.2 Why Least Median Squares Solution Does Not Exist

Recall that

- the least median squares estimator of linear regression is the value satisfying

$$\min_{\hat{\theta}} \text{med}_i r_i(\hat{\theta})^2$$

where $r_i(\hat{\theta}) = y_i - \boldsymbol{x}_i\hat{\theta}$;

- the maximum likelihood estimator is defined by

$$\min_{\hat{\theta}} \sum_{i=1}^{n} r_{Di}(\hat{\theta})^2,$$

where r_{Di} are the deviance residuals;

- the minimum Pearson chi-squares estimator is defined by

$$\min_{\hat{\theta}} \sum_{i=1}^{n} r_{Pi}(\hat{\theta})^2,$$

where r_{Pi} are the Pearson residuals.

Now when thinking of a high breakdown point estimator of logistic regression, one may think there should exist a solution satisfying

$$\min_{\hat{\theta}} \text{med}_i r_{.i}(\hat{\theta})^2, \tag{7.2.1}$$

where $r_{.i}$ is r_{Di} or r_{Pi}. However, using (7.2.1) as an extension to the least median squares estimator in linear regression fails.

Proposition 7.1 *No finite optimum solution to the objective function* $\text{med}_i r_{.i}(\hat{\theta})^2$ *exists.*

This can be illustrated by the example of a simple logistic regression with fixed intercept α. Suppose one has a sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where y_i is the binary response whose probability of success $(y_i = 1)$ is assumed to be

$$p(x_i) = \frac{1}{1 + \exp(-(\alpha + \beta x_i))},$$

where $x_i \neq 0$ for $i = 1, \ldots, n$. Now let n_{00} be the number of the samples whose $x < 0$ and $y = 0$; n_{01} be the number of the samples whose $x < 0$ and $y = 1$; n_{10} be the number of the samples whose $x > 0$ and $y = 0$; and n_{11} be the number of the samples whose $x > 0$ and $y = 1$. Since

$$n_{00} + n_{01} + n_{10} + n_{11} = n,$$

we know one of the following inequalities is true,

$$n_{00} + n_{11} \geq \frac{n}{2}$$

or

$$n_{01} + n_{10} \geq \frac{n}{2}.$$

Without loss of generality, we assume $n_{00} + n_{11} \geq n/2$. Now as $\beta \to +\infty$, because

$$p(x) \to 0 \text{ for } x < 0$$

and

$$p(x) \to 1 \text{ for } x > 0,$$

there are over $n/2$ data points whose residuals (deviance or Pearson) are arbitrarily close to 0. Therefore the objective function, the median of residual squares, can be arbitrarily close to zero. And there will be no finite β achieving the minimum of the objective function.

7.3 A Global Robust Estimator via Smoothing

As we all know, the global robust estimator for grouped data in the last chapter does not work for the case of $m < 5$. The situation is even worse when $m = 1$ because every observation is the extreme response data, where the empirical logistic transformation is ill-defined. However, if one can obtain a reasonable estimate of the probability p_i based on m local observations, the global robust estimator proposed in last chapter can be extended to the case of ungrouped data. A local estimate of the p_i can be determined by a locally weighted smoothing technique.

7.3.1 The Locally Weighted Smoother and the Model

A large number of local smoothers have been proposed and studied in the last fifteen years. Without doubt, the topic of how to choose a suitable smoother for binary response deserves another dissertation. For a review of local smoothers, see Hastie and Tibshirani (1992); for the implementation of local smoothers, see Chambers and Hastie (1992).

Here, we follow the smoothing process proposed by Fowlkes (1987), and extend it to the case of multiple explanatory variables. Suppose we have samples $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i$ is a $p-1$ row vector with continuous elements; y_i is a binary response. Now let N_i be the set of indices of the observations corresponding to the m nearest neighbors of $\boldsymbol{x}_i$ with respect to the Euclidean distance. Then the m-local neighborhood smoothed value, $\hat{p}_i$ is given by

$$\hat{p}_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \tag{7.3.1}$$

where the sum is over $j \in N_i$ and

$$w_{ij} = 1 - (d_{ij}/d_i)^3 \quad \text{for } j \in N_i \tag{7.3.2}$$

is a cubic weight function. Here d_{ij} is the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; and

$$d_i = \max_{j \in N_i} d_{ij}.$$

Theorem 7.1 *If for all* $j = 1, \ldots, n$,

$$p_j = 1/(1 + exp(-(\alpha + \boldsymbol{x}_j \theta))),$$

and the m *nearest neighbors of* $\boldsymbol{x}_i$ *are close enough to* $\boldsymbol{x}_i$, *then*

$$E(\hat{p}_i) \approx p_i + p_i(1 - p_i) \left[\frac{\sum_j w_{ij} \boldsymbol{x}_j}{\sum_j w_{ij}} - \boldsymbol{x}_i \right] \theta. \tag{7.3.3}$$

Proof: Let $\delta_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$. Now by Taylor's expansion, we have

$$\begin{aligned} p_j &= \frac{1}{1+\exp(-(\alpha+\boldsymbol{x}_j\theta))} \\ &= \frac{1}{1+\exp(-(\alpha+\boldsymbol{x}_i\theta+\delta_{ij}\theta))} \\ &\approx \frac{1}{1+\exp(-(\alpha+\boldsymbol{x}_i\theta))} + \delta_{ij}\frac{\exp(-(\alpha+\boldsymbol{x}_i\theta))}{(1+\exp(-(\alpha+\boldsymbol{x}_i\theta)))^2}\theta \\ &\approx p_i + p_i(1-p_i)\delta_{ij}\theta. \end{aligned}$$

Then

$$\begin{aligned} E(\hat{p}_i) &= \frac{\sum_j w_{ij}p_j}{\sum_j w_{ij}} \\ &\approx \frac{\sum_j w_{ij}p_i}{\sum_j w_{ij}} + \frac{\sum_j w_{ij}p_i(1-p_i)\delta_{ij}\theta}{\sum_j w_{ij}} \\ &\approx p_i + p_i(1-p_i)\frac{\sum_j w_{ij}\delta_{ij}}{\sum_j w_{ij}}\theta \\ &\approx p_i + p_i(1-p_i)\left[\frac{\sum_j w_{ij}\boldsymbol{x}_j}{\sum_j w_{ij}} - \boldsymbol{x}_i\right]\theta. \end{aligned}$$

□

Following the approach in last chapter, we examine the logit transformation of smoothed probability estimates.

Theorem 7.2 *With smoothed probability estimates given by (7.3.1), we have the logit linear model*

$$E[\log(\frac{\hat{p}_i}{1-\hat{p}_i})] = \tilde{\boldsymbol{x}}_i\theta, \tag{7.3.4}$$

where

$$\tilde{\boldsymbol{x}}_i = \frac{\sum_j w_{ij}\boldsymbol{x}_j}{\sum_j w_{ij}}. \tag{7.3.5}$$

Proof: It can be shown that if $\hat{p}_i$ and p_i are close, then by Taylor's expansion

$$\log(\frac{\hat{p}_i}{1-\hat{p}_i}) \approx \log(\frac{p_i}{1-p_i}) + \frac{\hat{p}_i - p_i}{p_i(1-p_i)}. \tag{7.3.6}$$

By Theorem 7.1, we have

$$E[\hat{p}_i] - p_i \approx p_i(1-p_i)(\tilde{x}_i - x_i)\theta.$$

Thus,

$$\begin{aligned} E[log(\frac{\hat{p}_i}{1-\hat{p}_i})] &\approx \log(\frac{p_i}{1-p_i}) + \frac{E[\hat{p}_i] - p_i}{p_i(1-p_i)} \\ &\approx x_i\theta + (\tilde{x}_i - x_i)\theta \\ &= \tilde{x}_i\theta. \end{aligned}$$

□

7.3.2 Global Robust Estimator

The estimator proposed here is an application of the empirical logistic least median squares estimator of the last chapter. The main problem in applying this estimator to ungrouped data is the failure of the empirical logistic transformation, and this can be overcome by the locally weighted smoothing technique. Specifically, one can use the locally weighted smoother discussed above to calculate , for each data point, the smoothed probability estimate whose empirical logit is well defined.

Suppose $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ are the sample data that are going to be fitted by logistic model

$$p_i = \frac{1}{1 + e^{-(\alpha + x_i\beta)}}.$$

Then the algorithm for obtaining a global robust estimator is described by the following,

1. **Locally weighted smoother.**

   For each $x_i, i = 1, \ldots, n$, find out D_i and calculate w_{ij} for $j \in D_i$ by (7.3.2). Then $\tilde{x}_i, \hat{p}_i$ are obtained by (7.3.1) and (7.3.5) respectively.

2. **Empirical Logit.**

The empirical logit for each (x_i, y_i) is defined by

$$z_i = \begin{cases} -\log(2\sum_j w_{ij} + 1) & \text{if } \hat{p}_i = 0; \\ \log(\frac{\hat{p}_i}{1-\hat{p}_i}) & \text{if } 0 < \hat{p}_i < 1; \\ \log(2\sum_j w_{ij} + 1) & \text{if } \hat{p}_i = 1. \end{cases} \tag{7.3.7}$$

3. **LMS estimator.**

The procedure described in 6.4 is used to calculate $\hat{\theta}^1$ of the model

$$E(Z) = \tilde{X}\theta$$

where

$$\tilde{X} = \begin{pmatrix} 1 & \tilde{x}_1 \\ \vdots & \vdots \\ 1 & \tilde{x}_n \end{pmatrix}$$

and $\theta = (\alpha, \beta^T)^T$.

4. **One step improvement.**

The same one step improvement formula is applied to the raw data with starting estimates $\hat{\theta}^1$ to obtain $\hat{\theta}^2$, that is,

$$\hat{\theta}^2 = \hat{\theta}^1 + (X^T M_c X)^{-1} X^T W \psi_c(r) \tag{7.3.8}$$

where

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

and the rest are defined as in section 6.5.2.

7.3.3 High Breakdown Point and Consistency

As in section 6.5.3, the breakdown in logistic regression for ungrouped data is defined as the case when all the probability estimates tend to be an extreme value, 0 or 1. Hence,

as suggested by section 6.5, we would expect that the breakdown point of $\hat{\theta}^1$ or $\hat{\theta}^2$, for given smoothed probability estimates, is

$$\epsilon_n^* = \frac{1}{2} - \frac{l}{n}, \tag{7.3.9}$$

where l is the maximum number of the smoothed probability estimates that are zeros or ones.

However, as we know, the smoothed probability estimates are not the raw data, and each estimate is not independent of its neighbors. The true breakdown point of the $\hat{\theta}^1$ or $\hat{\theta}^2$ would depend on the smoothing procedure as well as the distribution of the contaminated points. The followings are some typical cases of contamination that might happen,

- With moderate m, the smoothing procedure is robust to isolated outliers because of the weighted average.
- In the case of clustered outliers, the contamination would not or would barely spread out to the estimates of the other good data points because of the nearest neighborhood algorithm.
- When the contamination is everywhere in the space of R^k, the smoothing procedure yields biased estimates in many places, even perhaps for all the data points.

In the first two cases, the global robust estimator works properly to distinguish the outliers and good data points, while it fails for the third case in which it is almost impossible to separate the outliers from the good data points. Therefore, for the applicable cases, the breakdown point of $\hat{\theta}^1$ or $\hat{\theta}^2$ is $0.5 - l/n$.

Although the empirical logit obtained by the smoothed value are not independent of each other, the least squares estimator based on the empirical logit and $\tilde{X}$ is asymptotically consistent.

Theorem 7.3 *The least squares estimator of the model*

$$E(Z) = \tilde{X}\theta$$

is asymptotically consistent, where Z is defined by (7.3.7) with

$$\lim_{n\to\infty} \frac{m}{n} = 0$$

and $\tilde{X}$ is given by (7.3.5).

Proof: The least squares estimator of the model can be written as

$$\hat{\theta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T Z.$$

It is easy to show that $\hat{\theta}$ is an unbiased estimator, because

$$E(\hat{\theta}) = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T E(Z) = \theta.$$

Now we let $V = var(Z)$. By the algorithm we suggested, each $\hat{p}_i$ is correlated with other $\hat{p}_j$'s of the order of m. Thus there are of the order of m nonzero elements in each row/column of V.

The variance of $\hat{\theta}$ can be written as

$$var(\hat{\theta}) = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T V\tilde{X}(\tilde{X}^T\tilde{X})^{-1}.$$

Since

$$\tilde{X}^T\tilde{X} = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{x}}_i,$$

each element of $\tilde{X}^T\tilde{X}$ is of the order n. And because

$$\tilde{X}^T V\tilde{X} = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i^T \sum_{j=1}^{n} v_{ij}\tilde{\boldsymbol{x}}_j,$$

each element of $\tilde{X}^T V \tilde{X}$ is of the order mn. Accordingly, each element of $var(\hat{\theta})$ is of the order $1/n \cdot mn \cdot 1/n$, i.e m/n. Hence we have

$$\lim_{n \to \infty} var(\hat{\theta}) = 0;$$

that is, $\hat{\theta}$ is a consistent estimator of θ.

□

The consistency of the least squares estimator suggests the consistency of the $\hat{\theta}^2$. Simulation studies in the next chapter will demonstrate the properties of this proposed robust estimator.

Chapter 8

Some Simulation Results in Logistic Regression

Small simulation studies were performed to investigate the performance of the maximum likelihood estimator, robust estimators and the proposed global robust estimators ($\hat{\theta}^1$ and $\hat{\theta}^2$) under cases of non-contamination and contamination in logistic regression. The least absolute deviation estimator and bounded (leverage) influence estimator basically work for ungrouped data, but Beran's robust estimator is the only available robust estimator for both grouped and ungrouped data. All programs for the simulation studies were written by Splus and implemented on Sun Sparc 10 in Department of Biostatistics at MCV.

Random observations are mostly generated from a simple ($p = 2$) or multiple ($p = 3, 4$) logistic regression model

$$\log(\frac{p_i}{1-p_i}) = \alpha + \boldsymbol{x}_i\beta, \tag{8.0.1}$$

$$\log(\frac{p_i}{1-p_i}) = \alpha + \boldsymbol{x}_{1i}\beta_1 + \boldsymbol{x}_{2i}\beta_2, \tag{8.0.2}$$

or

$$\log(\frac{p_i}{1-p_i}) = \alpha + \boldsymbol{x}_{1i}\beta_1 + \boldsymbol{x}_{2i}\beta_2 + \boldsymbol{x}_{3i}\beta_3. \tag{8.0.3}$$

To study robustness of the estimators, we examine simulation results from the following three cases:

Table 8.1: Data Generation for Simple Logistic Regression

Case	X	Y
A	$x \sim N(0,1)$	$y \sim BIN(m, p(x))$
B	80%: $x \sim N(0,1)$ 20%: $x \sim N(-10,1)$	$y \sim BIN(m, p(x))$
C	80%: $x \sim N(0,1)$ 20%: $x \sim N(-10,1)$	$y \sim BIN(m, p(x))$ $y \sim$ Random Number

Table 8.2: Data Generation for Multiple Logistic Regression (p=3)

Case	X_1	X_2	Y
A	$x_1 \sim N(0,1)$	$x_2 \sim$ Uniform(0,1)	$y \sim BIN(m, p(x_1, x_2))$
B	80%: $x_1 \sim N(0,1)$ 20%: $x_1 \sim N(-10,1)$	$x_2 \sim$ Uniform(0,1)	$y \sim BIN(m, p(x_1, x_2))$
C	80%: $x_1 \sim N(0,1)$ 20%: $x_1 \sim N(-10,1)$	$x_2 \sim$ Uniform(0,1)	80%: $y \sim BIN(m, p(x_1, x_2))$ 20%: $y \sim$ Random Number

Table 8.3: Data Generation for Multiple Logistic Regression (p=4)

Case	X_1	X_2	X_3	Y
A	$x_1 \sim N(0,1)$	$x_2 \sim$ Uniform(0,1)	$x_3 \sim N(0,1)$	$y \sim BIN(m, p(x_1, x_2))$
B	80%: $x_1 \sim N(0,1)$ 20%: $x_1 \sim N(-10,1)$	$x_2 \sim$ Uniform(0,1)	$x_3 \sim N(0,1)$	$y \sim BIN(m, p(x_1, x_2))$
C	80%: $x_1 \sim N(0,1)$ 20%: $x_1 \sim N(-10,1)$	$x_2 \sim$ Uniform(0,1)	$x_3 \sim N(0,1)$	80%: $y \sim BIN(m, p(x_1, x_2))$ 20%: $y \sim$ Random Number

Case A: n random observations $(x_i, y_i), i = 1, \ldots, n$ are generated from the model (8.0.1), (8.0.2) or (8.0.3).

Case B: 80% of n random observations are generated as in case A; while the other 20% random observations are high leverage points in X space but their responses are correctly generated from the model.

Case C: 80% of n random observations are generated as in Case A; while the other 20% random observations are high leverage points in X space and their responses are random numbers from 0 through m.

The details of the cases for $p = 2$ and $p = 3, 4$ are summarized by table 8.1, table 8.2 and table 8.3.

To evaluate the performance of the estimators, we will compute the mean squares error matrix defined as

$$MSE(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)(\hat{\theta}_b - \theta)^T \tag{8.0.4}$$

$$= (\bar{\theta} - \theta)(\bar{\theta} - \theta)^T + \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})(\hat{\theta}_b - \bar{\theta})^T, \tag{8.0.5}$$

where B is the number of replications in the simulation and

$$\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b.$$

Definition 8.1 *The bias measure of the estimate is defined by*

$$Bias = (E\hat{\theta} - \theta)^T(E\hat{\theta} - \theta);$$

and the variance measure of the estimate is defined by the trace of the variance matrix of $\hat{\theta}$.

Hence, we estimate the bias and variance term of the estimates by replacing $E\hat{\theta}$ by $\bar{\theta}$ and the covariance matrix by the sample covariance matrix.

8.1 Simulation Studies for Grouped Data

For grouped data, we examine the cases of $m = 5$ and $m = 10$, with parameters $\alpha = -0.5$ and $\beta = 2$ in simple logistic regression; with parameters $\alpha = 1$, $\beta_1 = 3$, $\beta_2 = -2$ and parameters $\alpha = 1$, $\beta_1 = 3$, $\beta_2 = -2$, $\beta_3 = -0.5$ in multiple logistic regression of $p = 3$ and $p = 4$ respectively. The data of sample size of $n = 50$ are generated according to m and parameters for cases A-C, then the maximum likelihood estimator (MLE), Beran's robust estimator (Beran) and the proposed global robust estimators ($\hat{\theta}^1$ and $\hat{\theta}^2$) are calculated for the generated data. The same procedure is repeated $B = 500$ times. Table 8.4, 8.6 and 8.8 show the mean and standard deviation of all the estimates for all combinations. Moreover, table 8.5, 8.7 and 8.9 list the bias and variance measure of each estimate.

Generally speaking, for the ideal case (case A), the biases of the MLE, Beran's robust estimates and the proposed global robust estimates $\hat{\theta}^2$ when $m = 10$ are ignorable, while the bias of $\hat{\theta}^2$ when $m = 5$ is noticeable. For case B, in which $\boldsymbol{x}_2$ has 20% outlying points, the MLE, Beran's robust estimates and $\hat{\theta}^2$ with $m = 10$ behave much the same as in the case A, while the bias of $\hat{\theta}^2$ with $m = 5$ increases. For the case C, in which 20% contamination exists, the MLE wildly diverges, while the Beran's robust estimates are quite stable except for when $p = 4$ and $m = 10$. On the other hand, $\hat{\theta}^2$ behaves considerable better than all the other estimates when $m = 10$. Fig.8.1 shows the Bias for MLEs, Beran's estimates and $\hat{\theta}^2$ (in the figure $\hat{\theta}^2$ is named as "One" because it is the result of one step improvement procedure).

All $\hat{\theta}^1$ appear to be biased compared to the other estimates. The effect of the one step improvement can be seen from table 8.4 through table 8.9. Hence the one step improvement is certainly crucial for us in obtaining an unbiased estimator. Moreover, since almost all $\hat{\theta}^2$ for $m = 5$ have the worst biases, a moderate number of replications—m, say $m \geq 10$ is

Table 8.4: Means (S.D) of estimates for simple regression (p=2)

Case	m	Estimator	$\alpha = -0.5$	$\beta = 2$
A	5	MLE	-0.510 (0.170)	2.024 (0.270)
		Beran	-0.517 (0.185)	2.045 (0.326)
		$\hat{\theta}^1$	-0.398 (0.156)	1.548 (0.227)
		$\hat{\theta}^2$	-0.484 (0.167)	1.901 (0.244)
	10	MLE	-0.502 (0.117)	2.021 (0.185)
		Beran	-0.508 (0.126)	2.033 (0.220)
		$\hat{\theta}^1$	-0.440 (0.133)	1.788 (0.232)
		$\hat{\theta}^2$	-0.496 (0.125)	1.996 (0.211)
B	5	MLE	-0.418 (0.177)	1.926 (0.278)
		Beran	-0.437 (0.194)	1.949 (0.319)
		$\hat{\theta}^1$	-0.334 (0.171)	1.526 (0.282)
		$\hat{\theta}^2$	-0.408 (0.176)	1.838 (0.301)
	10	MLE	-0.493 (0.133)	1.992 (0.205)
		Beran	-0.501 (0.145)	2.019 (0.251)
		$\hat{\theta}^1$	-0.443 (0.147)	1.765 (0.256)
		$\hat{\theta}^2$	-0.492 (0.142)	1.971 (0.235)
C	5	MLE	-	-
		Beran	-0.502 (0.206)	2.072 (0.306)
		$\hat{\theta}^1$	-0.386 (0.188)	1.499 (0.307)
		$\hat{\theta}^2$	-0.467 (0.195)	1.864 (0.358)
	10	MLE	-	-
		Beran	-0.511 (0.145)	2.037 (0.253)
		$\hat{\theta}^1$	-0.451 (0.149)	1.775 (0.275)
		$\hat{\theta}^2$	-0.501 (0.143)	1.985 (0.251)

Table 8.5: Measures of estimators in simple regression (p=2)

m	Estimator	Case A		Case B		Case C	
		Bias	Var	Bias	Var	Bias	Var
5	MLE	0.0007	0.102	0.012	0.108	-	-
	Beran	0.0023	0.141	0.007	0.139	0.0053	0.173
	$\hat{\theta}^1$	0.215	0.076	0.252	0.109	0.263	0.130
	$\hat{\theta}^2$	0.010	0.088	0.035	0.122	0.0196	0.166
10	MLE	0.00046	0.048	0.0001	0.060	-	-
	Beran	0.0011	0.065	0.0004	0.084	0.0015	0.085
	$\hat{\theta}^1$	0.0487	0.072	0.0585	0.089	0.053	0.098
	$\hat{\theta}^2$	0.00003	0.060	0.0009	0.075	0.0002	0.084

Table 8.6: Means (S.D) of estimates for multiple regression (p=3)

Case	m	Estimator	$\alpha = 1$	$\beta_1 = 3$	$\beta_2 = -2$
A	5	MLE	1.012 (0.421)	3.081 (0.397)	-2.019 (0.748)
		Beran	1.047 (0.505)	3.177 (0.581)	-2.089 (0.896)
		$\hat{\theta}^1$	0.691 (0.380)	2.130 (0.417)	-1.381 (0.672)
		$\hat{\theta}^2$	0.882 (0.398)	2.698 (0.413)	-1.762 (0.698)
	10	MLE	1.023 (0.300)	3.055 (0.309)	-2.038 (0.533)
		Beran	1.028 (0.337)	3.086 (0.387)	-2.054 (0.605)
		$\hat{\theta}^1$	0.881 (0.342)	2.626 (0.368)	-1.751 (0.592)
		$\hat{\theta}^2$	0.992 (0.320)	2.964 (0.341)	-1.976 (0.562)
B	5	MLE	1.079 (0.436)	2.947 (0.419)	-2.014 (0.788)
		Beran	1.084 (0.491)	3.033 (0.604)	-2.051 (0.883)
		$\hat{\theta}^1$	0.758 (0.411)	2.024 (0.394)	-1.414 (0.742)
		$\hat{\theta}^2$	0.940 (0.417)	2.563 (0.427)	-1.764 (0.755)
	10	MLE	1.024 (0.312)	3.009 (0.316)	-1.998 (0.528)
		Beran	1.053 (0.363)	3.069 (0.423)	-2.067 (0.627)
		$\hat{\theta}^1$	0.887 (0.366)	2.591 (0.378)	-1.736 (0.640)
		$\hat{\theta}^2$	0.996 (0.333)	2.924 (0.345)	-1.955 (0.576)
C	5	MLE	-	-	-
		Beran	1.088 (0.577)	3.267 (0.742)	-2.156 (1.017)
		$\hat{\theta}^1$	0.635 (0.395)	1.765 (0.394)	-1.217 (0.631)
		$\hat{\theta}^2$	0.846 (0.416)	2.385 (0.462)	-1.637 (0.710)
	10	MLE	-	-	-
		Beran	1.009 (0.352)	3.105 (0.438)	-2.023 (0.626)
		$\hat{\theta}^1$	0.869 (0.354)	2.570 (0.433)	-1.742 (0.639)
		$\hat{\theta}^2$	0.969 (0.326)	2.923 (0.376)	-1.941 (0.574)

Table 8.7: Measures of estimators in multiple regression (p=3)

m	Estimator	Case A		Case B		Case C	
		Bias	Var	Bias	Var	Bias	Var
5	MLE	0.007	0.894	0.0092	0.986	-	-
	Beran	0.0416	1.396	0.0107	1.386	0.114	2.106
	$\hat{\theta}^1$	1.237	0.770	1.354	0.875	2.272	0.709
	$\hat{\theta}^2$	0.162	0.815	0.250	0.925	0.533	0.810
10	MLE	0.005	0.469	0.0007	0.476	-	-
	Beran	0.011	0.629	0.0119	0.705	0.0116	0.708
	$\hat{\theta}^1$	0.216	0.602	0.250	0.686	0.268	0.720
	$\hat{\theta}^2$	0.002	0.534	0.0078	0.562	0.0103	0.577

Table 8.8: Means (S.D) of estimates for multiple regression (p=4)

Case	m	Estimator	$\alpha = 1$	$\beta_1 = 3$	$\beta_2 = -2$	$\beta_3 = -0.5$
A	5	MLE	1.023 (0.406)	3.044 (0.398)	-2.061 (0.726)	- 0.512 (0.121)
		Beran	1.109 (0.468)	3.120 (0.581)	-2.165 (0.865)	- 0.557 (0.154)
		$\hat{\theta}^1$	0.644 (0.393)	2.029 (0.361)	-1.340 (0.695)	-0.346 (0.123)
		$\hat{\theta}^2$	0.862 (0.378)	2.589 (0.358)	-1.742 (0.677)	-0.438 (0.115)
	10	MLE	1.016 (0.291)	3.059 (0.299)	-2.030 (0.506)	-0.505 (0.087)
		Beran	1.043 (0.331)	3.109 (0.392)	-2.071 (0.581)	-0.526 (0.107)
		$\hat{\theta}^1$	0.848 (0.346)	2.600 (0.361)	-1.691 (0.608)	-0.429 (0.108)
		$\hat{\theta}^2$	0.968 (0.308)	2.944 (0.337)	-1.932 (0.539)	-0.485 (0.097)
B	5	MLE	1.065 (0.437)	2.932 (0.416)	-1.925 (0.791)	-0.488 (0.133)
		Beran	1.091 (0.498)	3.035 (0.670)	-1.989 (0.910)	-0.538 (0.178)
		$\hat{\theta}^1$	0.747 (0.437)	1.947 (0.469)	-1.326 (0.816)	-0.336 (0.140)
		$\hat{\theta}^2$	0.917 (0.405)	2.467 (0.495)	-1.650 (0.756)	-0.420 (0.129)
	10	MLE	1.041 (0.331)	3.004 (0.319)	-2.046 (0.584)	-0.500 (0.099)
		Beran	1.075 (0.352)	3.062 (0.429)	-2.115 (0.640)	-0.521 (0.114)
		$\hat{\theta}^1$	0.870 (0.399)	2.566 (0.403)	-1.730 (0.687)	-0.429 (0.117)
		$\hat{\theta}^2$	0.986 (0.355)	2.903 (0.369)	-1.963 (0.614)	-0.484 (0.106)
C	5	MLE	-	-	-	-
		Beran	1.171 (0.633)	3.314 (0.825)	-2.280 (1.133)	-0.603 (0.227)
		$\hat{\theta}^1$	0.566 (0.410)	1.625 (0.426)	-1.141 (0.670)	-0.284 (0.108)
		$\hat{\theta}^2$	0.785 (0.432)	2.240 (0.535)	-1.568 (0.722)	-0.389 (0.116)
	10	MLE	-	-	-	-
		Beran	0.997 (2.494)	3.226 (0.807)	-1.666 (10.145)	-0.693 (2.526)
		$\hat{\theta}^1$	0.835 (0.436)	2.461 (0.501)	-1.633 (0.823)	-0.408 (0.145)
		$\hat{\theta}^2$	0.964 (0.388)	2.841 (0.436)	-1.898 (0.736)	-0.474 (0.122)

Table 8.9: Measure of estimators in multiple regression (p=4)

m	Estimator	Case A		Case B		Case C	
		Bias	Var	Bias	Var	Bias	Var
5	MLE	0.0063	0.865	0.0147	1.008	-	-
	Beran	0.0567	1.300	0.0111	1.556	0.2168	2.417
	$\hat{\theta}^1$	1.5140	0.783	1.6530	1.096	2.8625	0.811
	$\hat{\theta}^2$	0.2583	0.742	0.421	0.996	0.8219	1.008
10	MLE	0.0047	0.438	0.0038	0.563	-	-
	Beran	0.0195	0.612	0.0230	0.730	0.1993	116.166
	$\hat{\theta}^1$	0.2833	0.631	0.2833	0.807	0.4608	1.139
	$\hat{\theta}^2$	0.0090	0.508	0.0112	0.651	0.0379	0.897

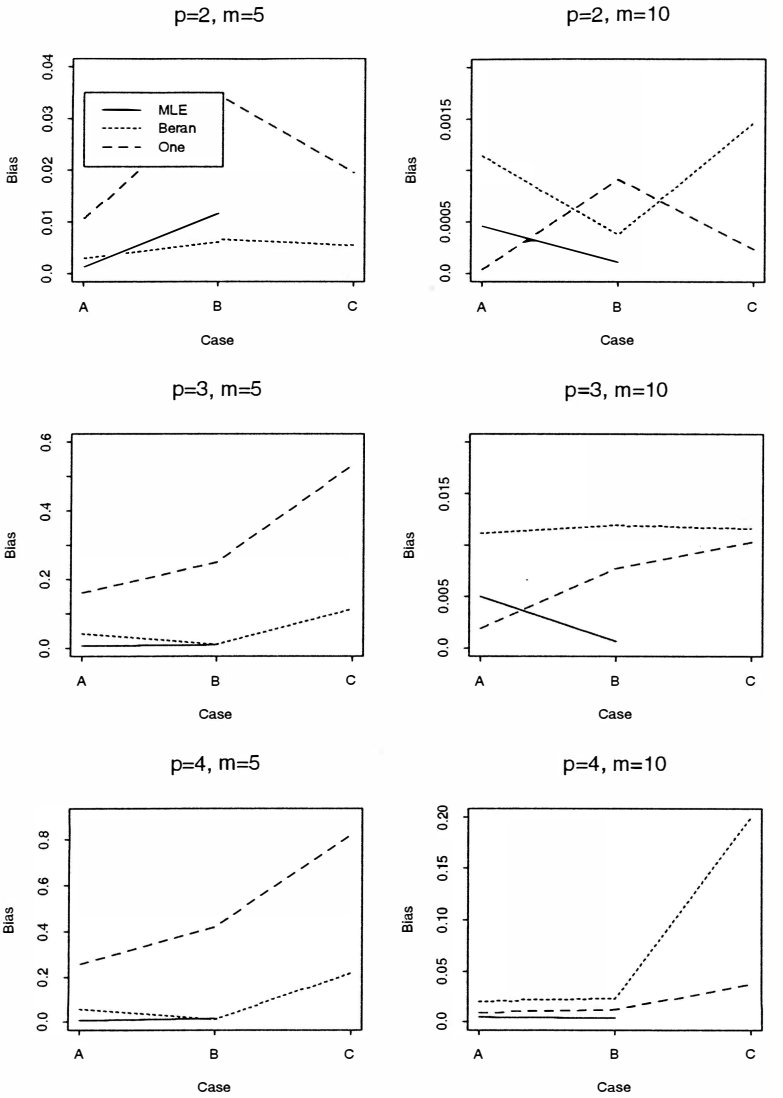

Figure 8.1: Bias of MLE, Beran and $\hat{\theta}^2$ For Grouped Data

another key to obtain an unbiased estimate. This also was the rationale in choosing $m = 10$ as the smoothing parameter for the robust procedure suggested in chapter 7.

8.2 Simulation Studies for Ungrouped Data

For the case of binary responses without replications, we examine the estimators of simple regression with parameters $\alpha = -0.5$ and $\beta = 2$; and multiple regression with parameters $\alpha = 1$, $\beta_1 = 3$, $\beta_2 = -2$ and $\alpha = 1$, $\beta_1 = 3$, $\beta_2 = -2$, $\beta_3 = -0.5$ for $p = 3$ and $p = 4$ respectively. Data of sample size $n = 100$ are generated for case A-C, then the maximum likelihood estimator (MLE), Beran's robust estimator (Beran), least absolute deviation estimator (LAD), bounded-leverage estimator (BL), bounded influence estimator (BI) and the proposed global robust estimators via smoothing ($\hat{\theta}^1$ and $\hat{\theta}^2$) are calculated for the generated data. The same procedure is repeated $B = 300$ times. Table 8.10, 8.12 and 8.14 show the mean and standard deviation of all the estimators for all the combinations. Table 8.11, 8.13 and 8.15 list the bias and variance measures of each estimate.

Notice that the MLEs are the best for case A and case B, whereas they diverge for case C. In terms of bias of the estimates, all the estimators including Beran, LAD, BL, BI and $\hat{\theta}^2$ behave quite well for case A and case B. And, as we anticipated, the variances of Beran and LAD estimators are large relative to the other estimators. For the case C, in which we can examine the robustness of the estimators, BL and BI estimates approach zero and thus show large biases. Beran and LAD estimators are quite stable if the procedures converge. Table 8.16 lists the number of occurrences of divergence in 300 replications. A large number of divergences happens in the Beran and LAD estimation procedures, especially when $p = 4$ for case C, 17.7% and 57.7% of estimates diverge for the Beran and LAD procedure respectively. One may imagine as the dimension of parameters increases,

Table 8.10: Means (S.D) of estimates for simple regression (p=2)

Case	Estimator	$\alpha = -0.5$	$\beta = 2$
A	MLE	-0.506 (0.292)	2.067 (0.458)
	Beran	-0.406 (2.248)	2.055 (1.674)
	LAD	-0.511 (0.307)	2.086 (0.511)
	BL	-0.506 (0.303)	2.072 (0.491)
	BI	-0.507 (0.303)	2.079 (0.491)
	$\hat{\theta}^1$	-0.428 (0.308)	1.695 (0.485)
	$\hat{\theta}^2$	-0.533 (0.335)	2.158 (0.530)
B	MLE	-0.468 (0.264)	1.999 (0.446)
	Beran	-0.515 (0.313)	2.090 (0.672)
	LAD	-0.482 (0.278)	2.008 (0.497)
	BL	-0.476 (0.274)	1.993 (0.490)
	BI	-0.476 (0.274)	1.999 (0.487)
	$\hat{\theta}^1$	-0.381 (0.315)	1.634 (0.522)
	$\hat{\theta}^2$	-0.504 (0.319)	2.076 (0.573)
C	MLE	-	-
	Beran	-0.566 (0.410)	2.314 (1.020)
	LAD	-0.536 (0.327)	2.159 (0.642)
	BL	-0.288 (0.208)	0.747 (0.297)
	BI	-0.287 (0.209)	0.743 (0.320)
	$\hat{\theta}^1$	-0.413 (0.443)	1.405 (0.702)
	$\hat{\theta}^2$	-0.489 (0.343)	1.832 (0.879)

Table 8.11: Measures of estimators in simple regression (p=2)

Estimator	Case A		Case B		Case C	
	Bias	Var	Bias	Var	Bias	Var
MLE	0.0046	0.295	0.0010	0.269	-	-
Beran	0.0118	7.858	0.0084	0.550	0.1028	1.208
LAD	0.0075	0.355	0.0004	0.324	0.0266	0.520
BL	0.0053	0.333	0.0006	0.315	1.6162	0.132
BI	0.0063	0.333	0.0006	0.312	1.6250	0.146
$\hat{\theta}^1$	0.0983	0.330	0.1479	0.371	0.3620	0.690
$\hat{\theta}^2$	0.0260	0.393	0.0058	0.430	0.0285	0.890

Table 8.12: Means (S.D) of estimates for multiple regression (p=3)

Case	Estimator	$\alpha = 1$	$\beta_1 = 3$	$\beta_2 = -2$
A	MLE	1.125 (0.688)	3.269 (0.743)	-2.252 (1.212)
	Beran	1.207 (0.926)	3.480 (1.181)	-2.386 (1.531)
	LAD	1.153 (0.771)	3.354 (0.921)	-2.305 (1.393)
	BL	1.125 (0.708)	3.285 (0.795)	-2.255 (1.272)
	BI	1.127 (0.709)	3.292 (0.798)	-2.259 (1.274)
	$\hat{\theta}^1$	0.788 (0.599)	2.264 (0.549)	-1.582 (1.030)
	$\hat{\theta}^2$	1.035 (0.683)	3.000 (0.657)	-2.077 (1.183)
B	MLE	1.096 (0.835)	3.291 (0.949)	-2.163 (1.518)
	Beran	1.134 (1.135)	3.524 (1.540)	-2.289 (2.088)
	LAD	1.120 (0.945)	3.386 (1.251)	-2.222 (1.773)
	BL	1.089 (0.850)	3.301 (1.034)	-2.157 (1.566)
	BI	1.092 (0.854)	3.310 (1.042)	-2.162 (1.573)
	$\hat{\theta}^1$	0.643 (0.549)	1.823 (0.328)	-1.247 (0.993)
	$\hat{\theta}^2$	0.863 (0.633)	2.583 (0.442)	-1.685 (1.144)
C	MLE	-	-	-
	Beran	1.277 (1.241)	3.624 (1.611)	-2.481 (2.189)
	LAD	1.304 (1.099)	3.435 (1.507)	-2.459 (1.853)
	BL	0.645 (0.432)	0.783 (0.348)	-1.119 (0.755)
	BI	0.651 (0.432)	0.773 (0.370)	-1.123 (0.755)
	$\hat{\theta}^1$	0.643 (0.548)	1.717 (0.525)	-1.255 (0.957)
	$\hat{\theta}^2$	0.917 (0.638)	2.463 (0.749)	-1.794 (1.146)

Table 8.13: Measures of estimators in multiple regression (p=3)

Estimator	Case A		Case B		Case C	
	Bias	Var	Bias	Var	Bias	Var
MLE	0.152	2.496	0.1206	3.903	-	-
Beran	0.422	4.597	0.3763	8.020	0.6970	8.926
LAD	0.242	3.385	0.2126	5.600	0.492	6.911
BL	0.162	2.751	0.1232	4.244	5.8199	0.877
BI	0.168	2.762	0.1308	4.287	5.8489	0.894
$\hat{\theta}^1$	0.762	1.721	2.0789	1.395	2.3298	1.492
$\hat{\theta}^2$	0.0072	2.299	0.2919	1.904	0.3374	2.282

Table 8.14: Means (S.D) of estimates for multiple regression (p=4)

Case	Estimator	$\alpha = 1$	$\beta_1 = 3$	$\beta_2 = -2$	$\beta_3 = -0.5$
A	MLE	1.083 (0.725)	3.372 (0.799)	-2.160 (1.250)	-0.546 (0.370)
	Beran	1.288 (1.212)	3.820 (1.675)	-2.535 (1.967)	-0.634 (0.590)
	LAD	1.153 (0.837)	3.519 (1.047)	-2.298 (1.446)	-0.578 (0.451)
	BL	1.069 (0.709)	3.368 (0.810)	-2.137 (1.225)	-0.544 (0.375)
	BI	1.071 (0.710)	3.375 (0.813)	-2.140 (1.227)	-0.545 (0.375)
	$\hat{\theta}^1$	0.654 (0.517)	2.019 (0.397)	-1.321 (0.894)	-0.320 (0.295)
	$\hat{\theta}^2$	0.892 (0.591)	2.794 (0.530)	-1.791 (1.012)	-0.447 (0.321)
B	MLE	1.111 (0.882)	3.413 (0.871)	-2.279 (1.534)	-0.553 (0.417)
	Beran	1.202 (1.152)	3.744 (1.488)	-2.279 (1.838)	-0.654 (0.572)
	LAD	1.116 (0.986)	3.595 (1.240)	-2.204 (1.661)	-0.604 (0.500)
	BL	1.112 (0.898)	3.426 (0.910)	-2.293 (1.568)	-0.547 (0.422)
	BI	1.115 (0.901)	3.435 (0.914)	-2.299 (1.573)	-0.548 (0.422)
	$\hat{\theta}^1$	0.527 (0.519)	1.441 (0.297)	-1.075 (0.872)	-0.214 (0.288)
	$\hat{\theta}^2$	0.754 (0.600)	2.220 (0.348)	-1.552 (1.024)	-0.354 (0.273)
C	MLE	-	-	-	-
	Beran	1.144 (1.062)	3.898 (1.898)	-2.490 (2.116)	-0.681 (0.544)
	LAD	1.089 (1.064)	3.520 (1.528)	-2.320 (1.898)	-0.592 (0.440)
	BL	0.612 (0.483)	0.521 (0.429)	-0.997 (0.785)	-0.237 (0.217)
	BI	0.612 (0.481)	0.511 (0.439)	-0.993 (0.781)	-0.237 (0.218)
	$\hat{\theta}^1$	0.488 (0.515)	1.204 (0.620)	-0.959 (0.819)	-0.220 (0.237)
	$\hat{\theta}^2$	0.748 (0.602)	1.855 (0.903)	-1.421 (0.971)	-0.336 (0.263)

Table 8.15: Measures of estimators in multiple regression (p=4)

Estimator	Case A		Case B		Case C	
	Bias	Var	Bias	Var	Bias	Var
MLE	0.173	2.862	0.2635	4.063	-	-
Beran	1.059	8.281	0.696	7.246	1.100	9.504
LAD	0.388	4.093	0.419	5.516	0.389	7.264
BL	0.161	2.799	0.2821	4.271	7.3741	1.080
BI	0.167	2.811	0.2936	4.300	7.4302	1.082
$\hat{\theta}^1$	1.575	1.311	3.590	1.202	4.6514	1.377
$\hat{\theta}^2$	0.101	1.758	0.891	1.604	1.7365	2.190

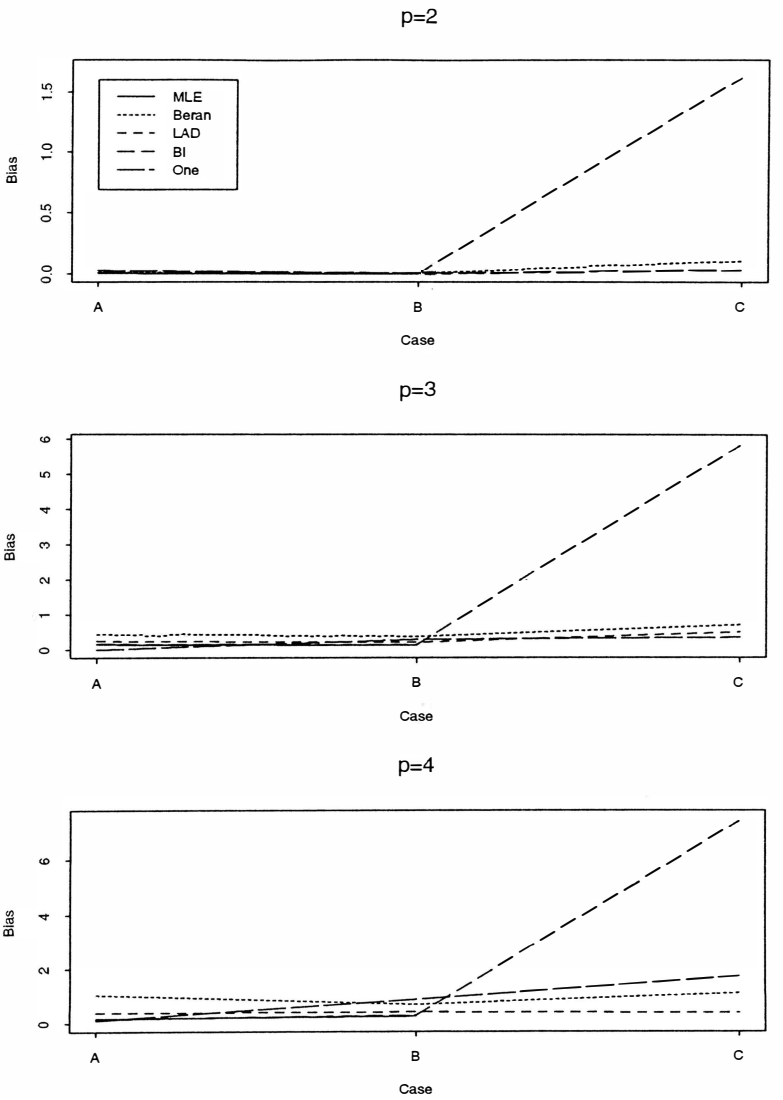

Figure 8.2: Bias of MLE, Beran, LAD, BI and $\hat{\theta}^2$ For Ungrouped Data

Table 8.16: Occurrence of Divergence of Beran/LAD in 300 replications

Case	Estimator	p		
		2	3	4
A	Beran	3	12	24
	LAD	0	0	1
B	Beran	5	19	47
	LAD	0	2	4
C	Beran	7	29	53
	LAD	18	136	173

divergence of the estimation procedure becomes a very important issue for both estimators. Whereas, there is no such problem with the proposed robust estimator, though the bias of the estimates are somewhat larger than that of the Beran or LAD estimator when $p = 4$ for case C. Fig.8.2 displays the Bias of MLE, Beran, LAD, BI and $\hat{\theta}^2$ (in the figure it is named as the One because of one step improvement procedure).

As in the last section, the effects of a one step improvement in the robust estimation procedure have been demonstrated by table 8.10 through 8.15.

8.3 Summary

The results of the simulation studies can be summarized by the following:

- The MLE, as anticipated, is the best both in terms of bias and variability of the estimates when the model is correctly specified.
- The Beran and LAD are robust to at least 20% contamination unless they diverge. And the divergence problem appears to be worse as the dimension of the parameters increases.
- The BL and BI break down at 20% contamination, in which the estimates approach zero, even though they have bounded leverage or influence.

- A moderate m, i.e. $m \geq 10$, is necessary for the suggested robust estimator to become an unbiased estimator.
- One step improvement in the suggested procedure is very effective in reducing the bias caused by the empirical logit transformation.
- $\hat{\theta}^2$ is robust to at least 20% clustered contamination.

However, because of limitations of resource and time, many issues remain unknown or unclear. For instance, how robust are the estimators under the varying mechanisms of the contamination? under a higher percentage, say $> 20\%$ of contamination? It is unknown what factors, the starting point, iteration times or the algorithm cause the considerable possibility of divergence of the Beran and LAD estimation procedure.

Chapter 9

Applications in Data Analysis

So far we have seen that the standardized influence matrix provides a general tool to evaluate the joint local influence of a parametric model, with applications in linear regression and logistic regression models. This chapter, by a few examples, will show that the diagnostic approach we suggested is effective in identifying multiple influential observations even with the presence of a masking effect, and the robust measure we proposed serves as a quantity to compare the robustness of the estimators with respect to joint local influence.

After introducing the data, we first examine diagnostic approaches to the maximum likelihood estimator (or least squares estimator in a linear model), then the robust measures of the estimators studied in chapter 5 and 6 are calculated. Discussion and interpretation are given for each example.

9.1 Examples in Linear Regression Model

9.1.1 Water Salinity Data

The data taken from Ruppert and Carroll (1980) are a typical example of a masking effect. It is a set of measurements of water salinity (i.e., salt concentration) and river discharge taken in North Carolina's Pamlico Sound. As Rousseeuw and Leroy (1987) suggested, we fit a linear model where salinity is regressed against salinity lagged by two weeks

(x_1), a linear time trend (x_2, the number of biweekly periods elapsed since the beginning of the spring season); and the volume of river discharge into the sound (x_3).

First we examine the diagnostic statistics given by the standardized residual, Cook's Distance, Cook's approach and our approach to local influence for the least squares estimator. Table 9.1 summarizes results for each statistics, where MC_2 stands for half of the maximum curvature. Fig.9.1-9.2 show results for each diagnostic procedure.

The standardized residual plot (Fig.9.1(a)) shows that case 16 is highly influential, and Cook's distance (Fig.9.1(b)) confirms it. Regarding the approach to local influence, half of the maximum curvature, or MC_2 of Cook's approach is 3.865, which is less than but close to the critical threshold $p = 4$ given by the np-guideline. Further examination of the corresponding eigenvector (Fig.9.2(a)) shows that adding point 16 might lead to some considerable changes to the values of the LS estimate. Therefore, point 16 has the potential to be an influential point. On the other hand, the largest eigenvalue of the standardized influence matrix of the least squares estimator with estimated parameters based on LMS, when divided by $n = 28$, is 1482.32. It is much larger than $p = 4$, indicating that there are some influential points in the data. They can be located by looking at the display of the eigenvector associated with the largest eigenvalue (Fig.9.2(b)). The display shows that adding point 16 and 5 would result in the parameter estimates substantially deviating from a global robust fit. Therefore, cases 16 and 5 are jointly influential points. In fact, Carroll and Ruppert (1985) indicated that cases 5 and 16 correspond to periods of very heavy discharge.

Huber's robust estimates and the Krasker-Welsch estimates are calculated with $c = 1.345$ and $a = 3.2$, respectively. With LS estimates as a starting point, the results of fitting (parameter estimates and the cases whose weights are much less than 1) and the

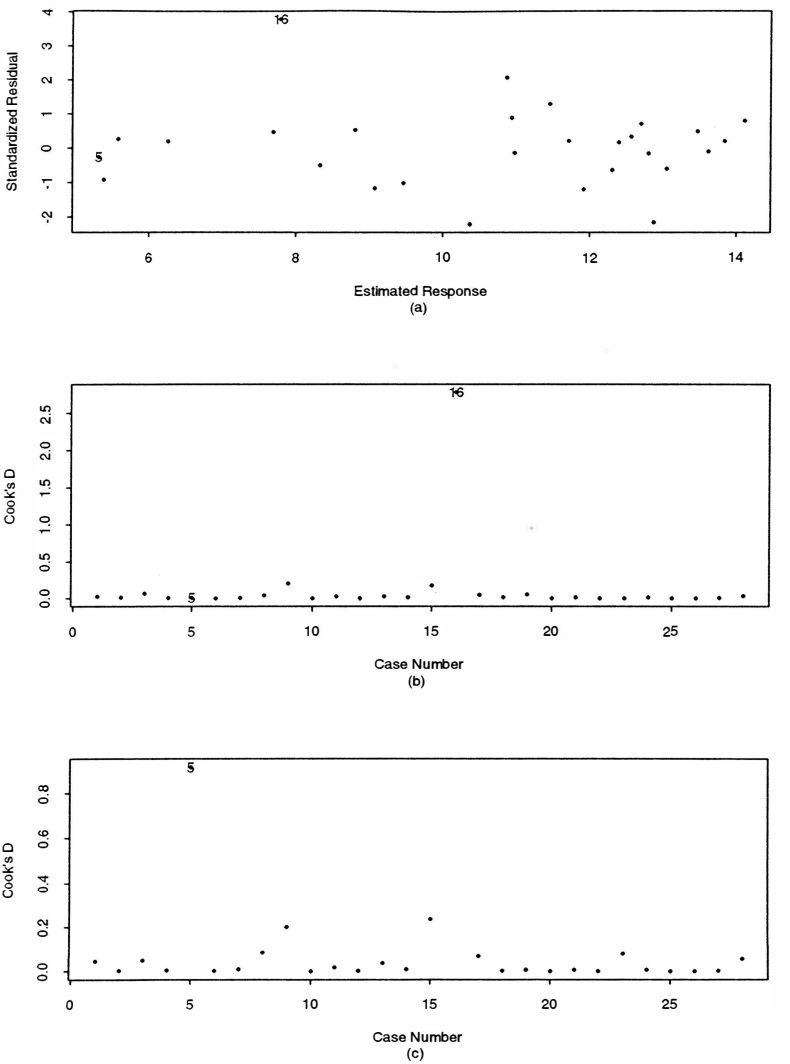

Figure 9.1: Deletion Diagnostics for Water Salinity Data, (a) Standardized residuals; (b) Cook's distance; (c) Cook's distance with case 16 deleted

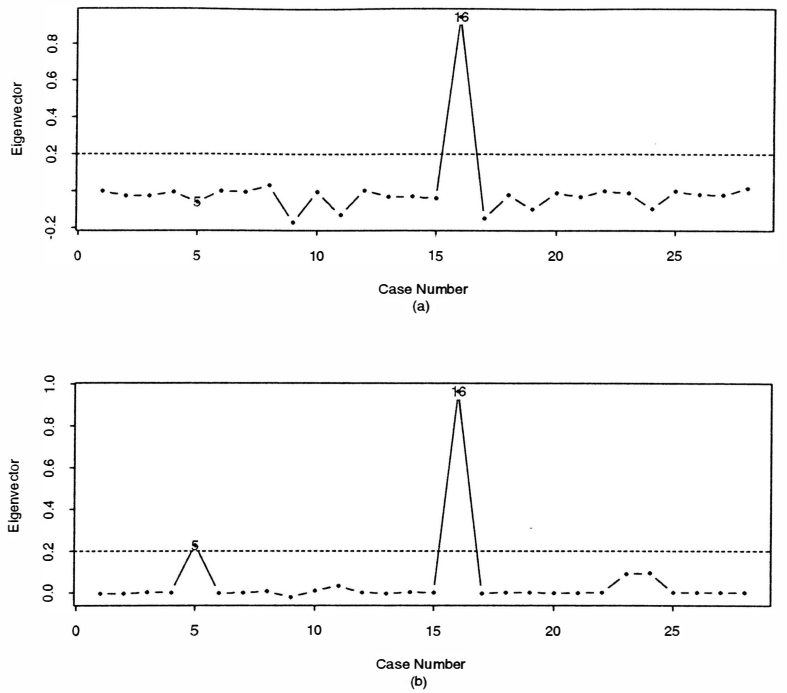

Figure 9.2: Local Influence Diagnostics for Water Salinity Data, (a) Cook's approach to local influence; (b) MSIF approach to local influence

Table 9.1: Diagnostics for Water Salinity Data

Method	Magnitude	Influential Points
Deletion	Maximum Value	
Standardized Residual	3.79	16, 9, 15, 17
Cook's Distance	2.78	16
Local Influence	MC_2	
Cook's Approach	3.865	16 (possibly)
Our Approach	1482.32	16, 5

Table 9.2: Comparison of Estimates of Water Salinity Data

Estimator	Parameter Estimates				Cases ($w.t \ll 1$)	MC_2
LS	9.59 (3.13)	0.78 (0.09)	-0.026 (0.16)	-0.30 (0.11)	None	1482.32
Huber	13.33 (2.40)	0.757 (0.066)	-0.094 (0.12)	-0.437 (0.082)	16, 15, 17	20.17
KW	16.95 (2.15)	0.722 (0.116)	-0.118 (0.137)	-0.581 (0.047)	16, 5, 9	2.47

measure of robustness, MC_2 of the influence graph are listed in table 9.2.

By looking at MC_2 in the table alone, we easily conclude that the Krasker-Welsch estimator is the most robust estimator, while the least squares estimator is the least. Moreover, we can consult with table 5.1 through 5.3 to see if the values of MC_2 are abnormal. Clearly, all the values are beyond the 95% quantile of the distribution of the largest eigenvalues of the **MSIF**s, suggesting that influential observations exist in data with respect to each estimate. However, the absolute values of MC_2 give a base to compare the stability of the estimates. Fig.9.3 further displays the eigenvectors associated with the largest eigenvalue for all three estimates. Fig.9.3(a) shows that small changes in cases 16 and 5 would lead to a substantial change in least squares estimates and inference. Since Huber's robust estimator has downweighted case 16 but not case 5, the influence caused by case 16 has been reduced while that caused by case 5 hasn't been. This is shown by Fig.9.3(b), in which the ratio between case 16 and case 5 become smaller, and small changes in cases 16, 5, 23, and 24 would lead to a large change in Huber's estimates. However, the change in Huber's estimates is much smaller than that in the least squares estimate. Finally, the final weights of cases 16, 5 and 9 in KW estimation are much less than 1. Thus, the Krasker-Welsch estimator succeeds in downweighting the influential points revealed above. Fig.9.3(c), in which cases 16, 5 are no longer the most influential points for KW-estimates, also shows

Table 9.3: Diagnostics for Hawkins Data

Method	Magnitude	Influential Points
Deletion	Maximum Absolute Value	
Standardized Residual	5.287	11-14
Cook's Distance	2.114	11-14
Local Influence	MC_2	
Cook's Approach	23.44	11-13
Our Approach	340.61	1-10

the effect of downweighting.

9.1.2 Hawkins-Bradu-Kass Data

The data set, generated by Hawkins, Bradu, and Kass (1984), consists of 75 observations with one response and three explanatory variables. The first 10 observations are bad leverage points and the next four points are good leverage points (their x_i are outlying, but the corresponding y_i fit the model quite well).

First, the diagnostic statistics given by the standardized residual, Cook's Distance, Cook's approach and our approach to local influence are calculated for the least squares estimator. Table 9.3 and Fig.9.4-9.5 display results. Basically, the first three approaches tell the same story, namely that cases 11-13 (14) are influential. MC_2 of Cook's approach is 23.44 (larger than $p = 4$), and the corresponding eigenvector is shown in Fig.9.5(a). It shows that adding points 11, 12, and 13 would lead to a substantial change of the values of the least squares estimator. Hence, we come to the wrong decision that the points 11, 12, and 13 are the most influential points. On the other hand, the largest eigenvalue of the standardized influence matrix, when divided by $n = 75$, is 340.61 and its associated eigenvector (Fig.9.5(b)) shows that a deletion of points 1-10 would make the value of the least squares estimator far away from the global robust fit. Therefore, points 1-10 are the influential points, and this is how the data was generated.

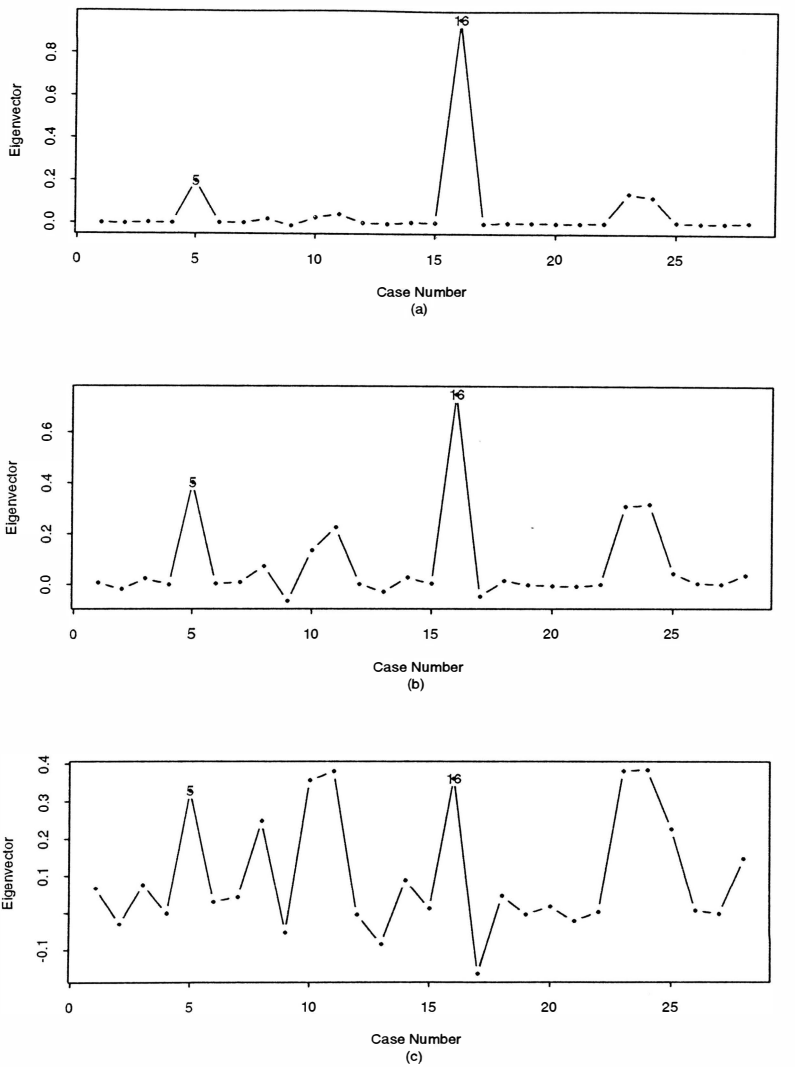

Figure 9.3: Eigenvectors of MSIF for LS, Huber and KW estimates of Water Salinity Data, (a) LS estimates; (b) Huber's estimates; (c) KW estimates

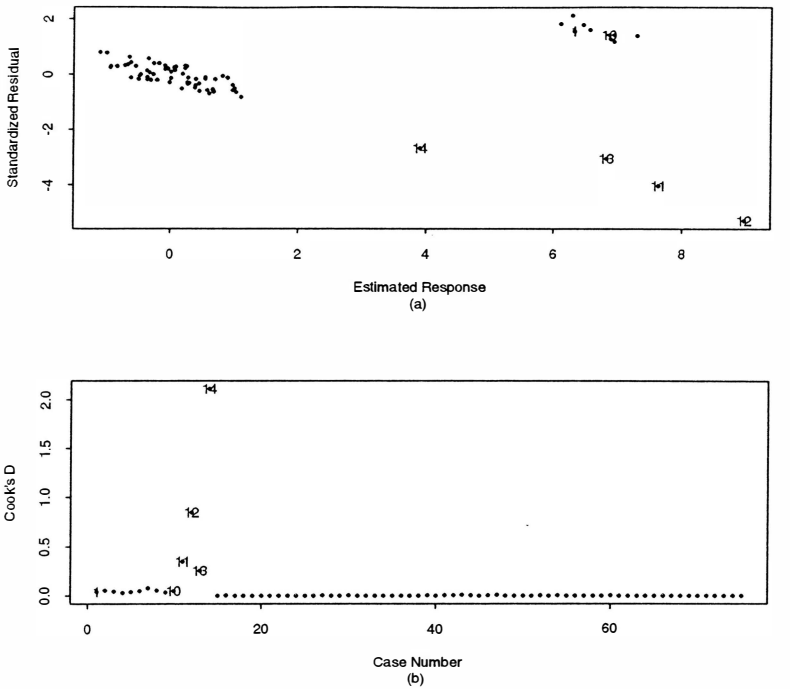

Figure 9.4: Deletion Diagnostics for Hawkins Data, (a) standardized residuals; (b) Cook's distance

Table 9.4: Comparison of Estimates of Hawkins Data

Estimator	Parameter Estimates				Cases ($w.t \ll 1$)	MC_2
LS	-0.39 (0.42)	0.24 (0.26)	-0.335 (0.16)	0.38 (0.13)	None	340.61
Huber	-0.780 (0.212)	0.166 (0.134)	0.011 (0.079)	0.273 (0.066)	11-14	4.72
KW	-0.778 (0.283)	0.164 (0.155)	0.060 (0.019)	0.228 (0.045)	11-14	1.93

Huber's robust estimates and the Krasker-Welsch estimates are calculated with $c = 1.345$ and $a = 3.2$, respectively. The results of fitting (parameter estimates and the cases whose weights are much less than 1) and the measure of robustness, MC_2 of the influence graph, are given in table 9.4.

As in the last example, the Krasker-Welsch estimator is the most robust one, while the LS estimator is the least, based on the value

of MC_2 in the table. By consulting table 5.1 through 5.3, one can conclude that the largest eigenvalues are abnormally large and that influential observations exist in the data for all three estimates. Fig.9.6 displays the eigenvectors associated with the largest eigenvalues for all three estimates. Fig.9.6(a) shows that small changes in cases 1-10 would lead to substantial changes in the least squares estimates and their inference. Since both Huber's estimates and the KW-estimates downweight cases 11-14, not cases 1-10, Fig.9.6(b) and Fig.9.6(c) suggest that small changes in cases 1-10 still result in considerable changes in both estimates, though the magnitude of the changes varies and is much smaller than that of the least squares estimates. Hence, this is an example of bounded influence estimates that break down due to a group of bad points. By examining the local influence of the estimates, one can find out if the bounded influence estimate is well enough protected against joint influence by a group of points.

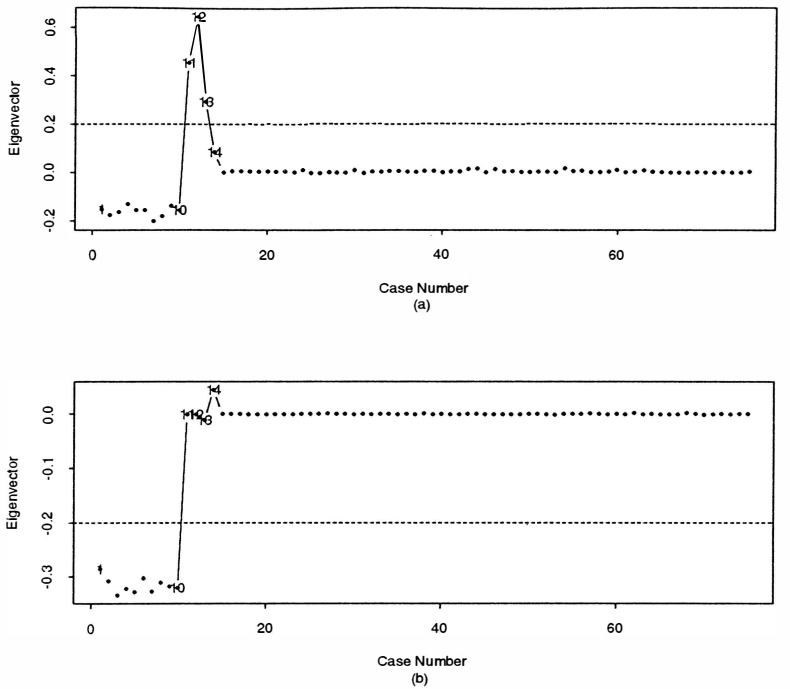

Figure 9.5: Local Influence Diagnostics for Hawkins Data, (a) Cook's approach to local influence; (b) MSIF approach to local influence

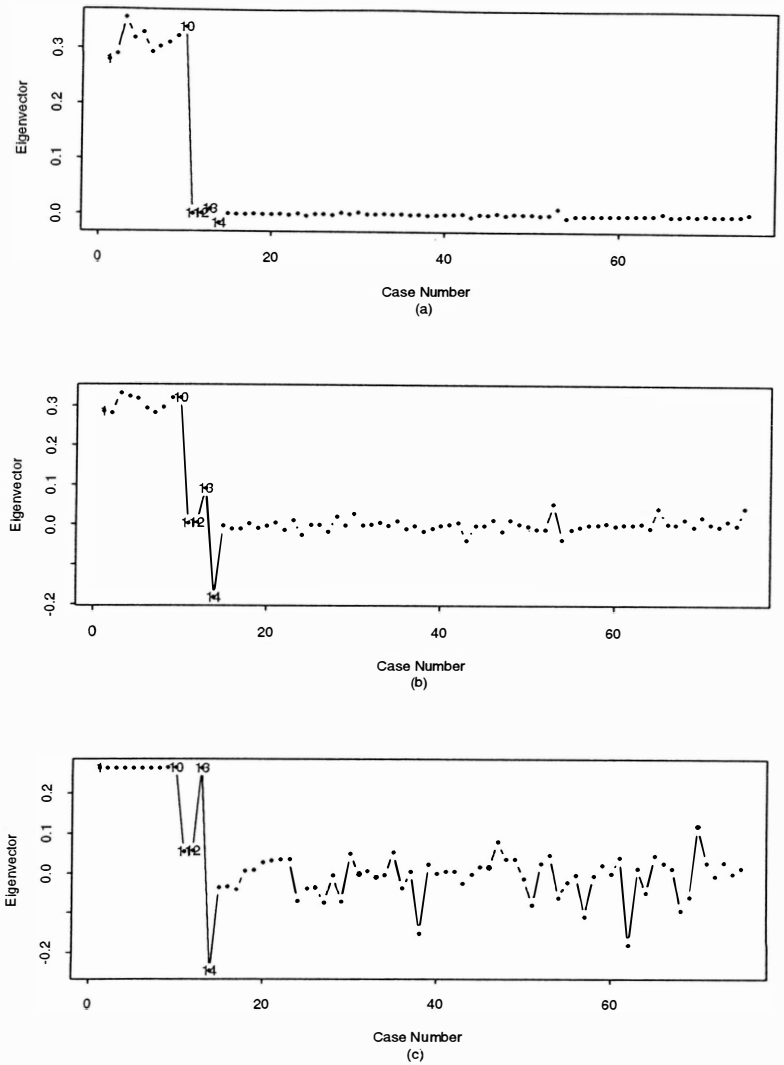

Figure 9.6: Eigenvectors of MSIF for LS, Huber and KW Estimates of Hawkins Data, (a) LS estimates; (b) Huber's estimates; (c) KW estimates

9.2 Examples in Logistic Regression Model

9.2.1 Vaso-Constriction Data

To illustrate the **MSIF** approach in a logistic regression context, we use data from Finney (1947) in which there are $n = 39$ cases. A binary response is defined in terms of the presence (1) or absence (0) of vaso-constriction of the skin of the digits after the inspiration of air. A logistic regression related this response to the volume of air, VOL, and the inspiration rate, RATE. Since this vaso-constriction data have been extensively studied by various researchers including Pregibon (1981, 1982), Copas (1988) and Morgenthaler (1992), they will be used to examine how effective is the proposed procedure.

The model is given as the following,

$$\text{logit}(p) = \beta_0 + \beta_1 \log(VOL) + \beta_2 \log(RATE). \tag{9.2.1}$$

First we examine diagnostic statistics for a maximum likelihood estimator. Table 9.5 and Fig.9.7-9.8 show results based on Cook's Distance, Cook's approach and our proposed approach to local influence. In fact all three approaches reveal the same influential points 4 and 18, which is consistent with the conclusions made by Pregibon (1981, 1982), Copas (1988), etc.. The difference here is that our MSIF approach gives a strong indication of being influenced by cases 4 and 18; while the deletion approach and Cook's approach to local influence result in weaker indications of influence.

The robust estimators discussed in chapter 6 and 7 are calculated and the estimates and their standard error (where applicable) are listed in table 9.6. Moreover, given cases 4 and 18 are influential, the MLE with cases 4 and 18 deleted (last entry of table 9.6) is also calculated to compare with the previous estimator. Having looked at table 9.6, one can find that the estimates by the LAD and the Beran estimators and the MLE with cases 4, 18

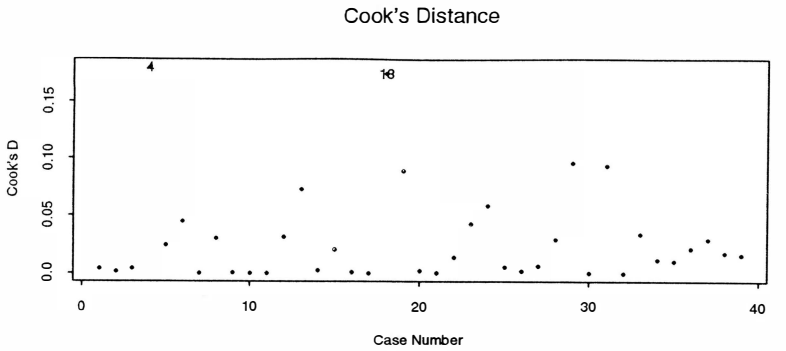

Figure 9.7: Deletion Diagnostics for Vaso-Constriction Data

Table 9.5: Diagnostics for Vaso-Constriction Data

Method	Magnitude	Influential Points
Deletion	Maximum Absolute Value	
Cook's Distance	0.180	4, 18
Local Influence	MC_2	
Cook's Approach	2.176	4, 18 (possibly)
Our Approach*	6.103	4, 18
Our Approach**	163.32	4, 18
Our Approach***	319.33	4, 18

* with $\hat{\theta}^2$ replacing the true parameters
** with LAD estimates replacing the true parameters
*** with MLE* replacing the true parameters

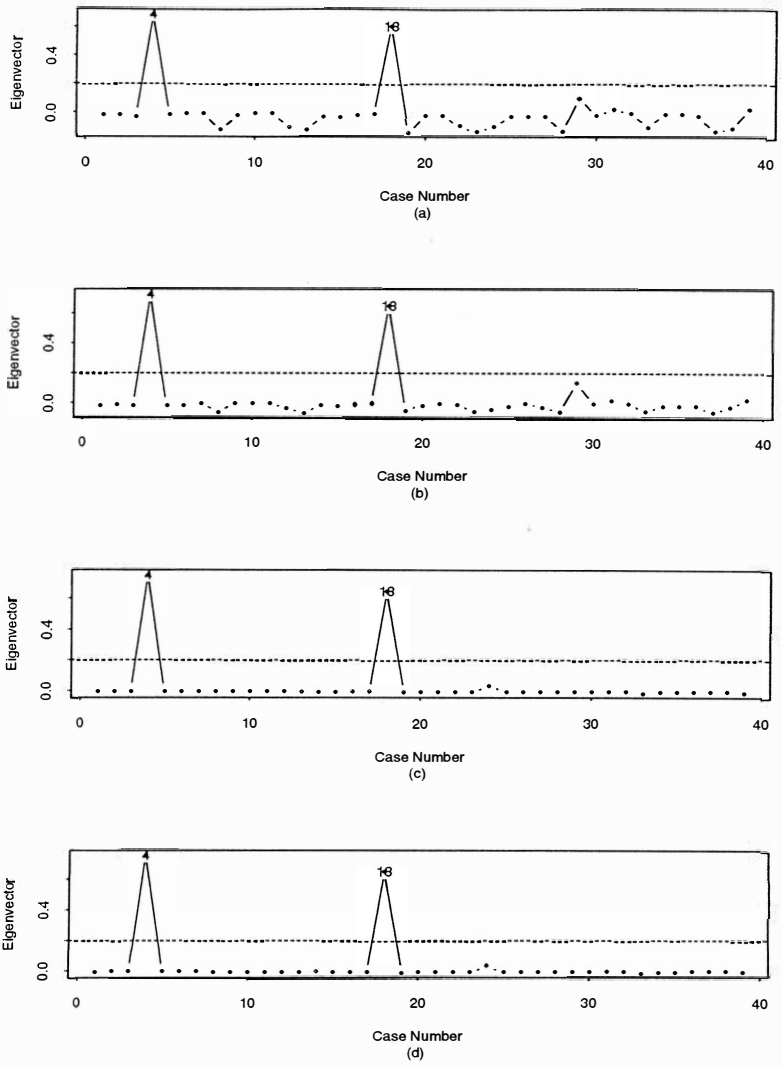

Figure 9.8: Local Influence Diagnostics for Vaso-Constriction Data, (a) Cook's approach; MSIF approach with true parameters replaced by (b) $\hat{\theta}^2$, (c) LAD estimates and (d) MLE*

deleted (MLE*), which are similar, depart substantially from the MLE estimates; while the estimates by the BL, BI estimators and $\hat{\theta}^2$ are somewhere between the two extremes. This phenomenon suggest that the first group that includes the LAD and Beran estimator and MLE*, is the most robust estimator group; while the second group that includes estimators BL, BI and $\hat{\theta}^2$, is less robust than the first group but more robust than the last group, comprised of MLE estimator. Actually, the following comparisons of the estimators by MC_2 of the **MSIF** approach, given by table 9.7, explain the phenomenon.

In table 9.7, MC_2 in the first column is obtained by using $\hat{\theta}^2$ to replace the true parameters in the formulation of the **MSIF**, in the second column by using LAD estimates and in the third column by using the MLE*. From the first column, the maximum curvature, 1.112 and 1.165 for Beran and LAD estimates (close to 1) suggests no cases would influence the estimates much. However, MC_2 for the BI estimates, 2.087, suggests considerable influence may happen on BI estimates by the data and the BI estimator is less robust than the Beran and LAD estimators. Finally, as revealed by the diagnostics, the MLE estimator is very much influenced by a small portion of the data and it is the least robust estimator. In Fig.9.9(a)-(b), the corresponding eigenvector plots locate cases 4 and 18 as the influential points for MLE and BI estimates. When the LAD estimates are chosen to replace the true parameters, a larger value of MC_2 for the MLE estimates supports the existence of influential points. However, MC_2s for the Beran, LAD and BI estimates are close to 2 and the difference in robustness between the Beran, LAD and BI estimates is no longer apparent. When the true parameters are replaced by the MLE*, the BI estimates appear to be most robust. In Fig.9.10-9.11, displays of the eigenvectors of **MSIF** approach with the true parameters replaced by LAD and MLE* are consistent with Fig.9.9, especially for panels (a) and (b).

Table 9.6: Estimates of Vaso-Constriction Data

Estimator	Intercept	$\log(VOL)$	$\log(RATE)$
MLE	-2.875	5.179	4.562
	(1.32)	(1.87)	(1.84)
Beran	-22.997	37.966	30.055
	(17.04)	(20.58)	(21.84)
LAD	-21.278	34.414	27.722
	(13.23)	(22.06)	(16.82)
BL	-5.427	8.596	7.679
	(2.38)	(3.59)	(3.19)
BI	-5.482	8.689	7.751
	(2.37)	(3.57)	(3.19)
$\hat{\theta}^1$	-4.222	4.997	5.751
$\hat{\theta}^2$	-5.636	7.771	7.615
MLE*	-24.581	39.550	31.935
	(14.01)	(23.22)	(17.74)

* Maximum likelihood estimate with cases 4 and 18 deleted

Unlike linear regression, the distribution of the largest eigenvalue in logistic regression depends on the parameters and subsequently on the selection of the parameter estimates to replace the true parameters. There seems no clear way of comparing the values of MC_2 between two different replacements of the true parameters. Hence, it appears that the robust measure, MC_2, depends on the data as well as the parameters. Quantifying the robustness of the estimator in terms of local influence may not be possible and the values may vary.

9.2.2 Low Birth Weight Data

The data, taken from Hosmer and Lemeshow (1989) are for a study of risk factors associated with low infant birth weight. The data were collected at Baystate Medical center, Springfield, Massachusetts during 1986. The variables identified in table 9.8 have been shown to be associated with low birth weight in obstetrical literature. The goal of the study was to ascertain if these variables were important in the population being served by

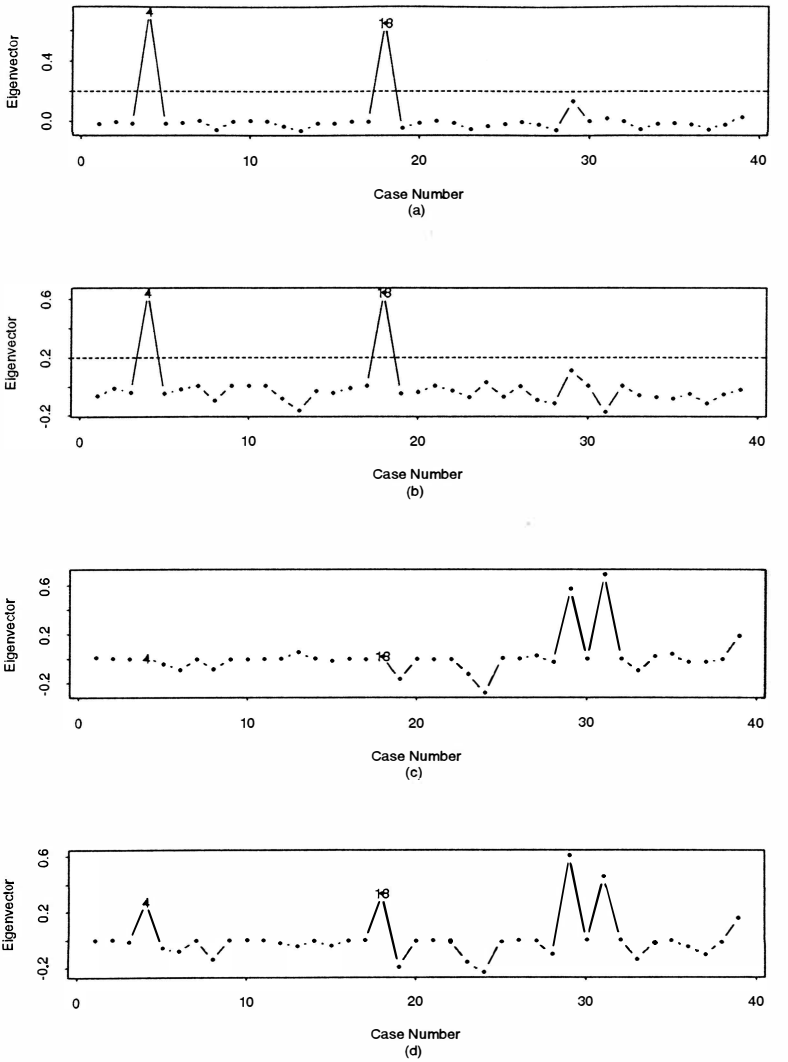

Figure 9.9: Eigenvectors of **MSIF** with $\hat{\theta}^2$ for Vaso-Constriction Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

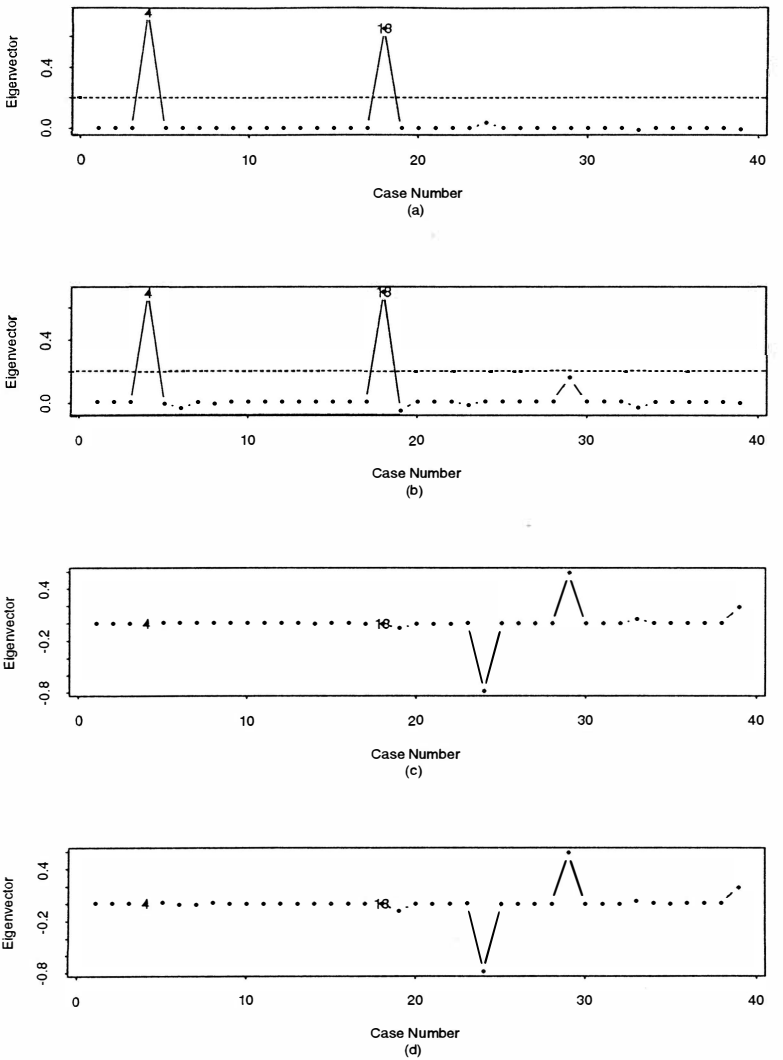

Figure 9.10: Eigenvectors of **MSIF** with LAD estimates for Vaso-Constriction Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

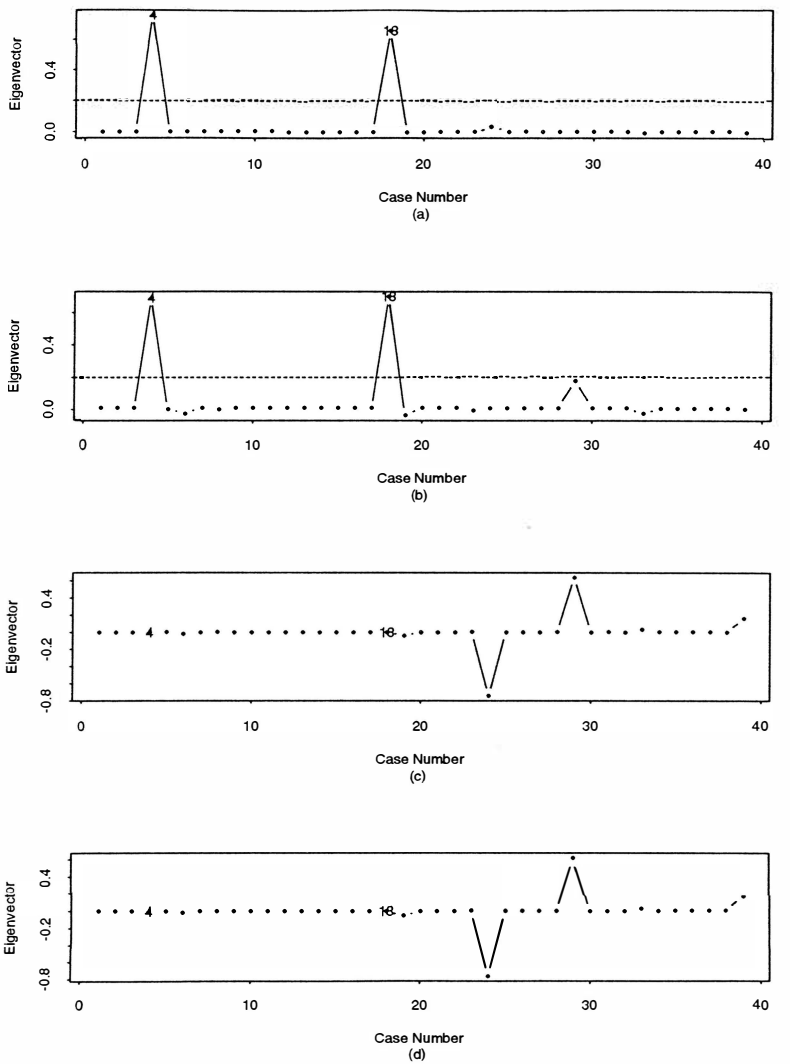

Figure 9.11: Eigenvectors of **MSIF** with MLE* for Vaso-Constriction Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

Table 9.7: Comparison of Estimates of Vaso-Constriction Data

Estimator	MC_2		
	$\tilde{\theta}^2$	LAD	MLE*
MLE	6.103	163.32	319.33
Beran	1.112	1.931	2.242
LAD	1.165	1.852	2.178
BI	2.087	1.813	1.816

the medical center where the data were collected.

Based on a detailed analysis, Hosmer and Lemeshow concluded that there were no associations between the AGE, FTV and low birth weight in this data. Hence we shall examine the model given by the following,

$$\begin{aligned} \text{logit}(p) &= \beta_0 + \beta_1 LWT + \beta_2 RACE1 + \beta_3 RACE2 + \beta_4 SMOKE \\ &\quad + \beta_5 PTL + \beta_6 HT + \beta_7 UI \end{aligned} \tag{9.2.2}$$

where

$$RACE1 = \begin{cases} 1 & RACE = 1 \\ 0 & \text{otherwise} \end{cases},$$

$$RACE2 = \begin{cases} 1 & RACE = 2 \\ 0 & \text{otherwise} \end{cases}.$$

As we did for the last example, we first examine diagnostic statistics for the maximum likelihood estimator. Table 9.9 and Fig.9.12-9.13 show results based on Cook's distance, Cook's approach and our proposed approach to local influence. From Fig.9.12(a), Cook's Distance suggests that cases 13 and 148 are relatively influential. However, unlike the previous example, after we delete case 13, case 51, which was not influential at first, emerges as the most influential point (see Fig.9.12(b), the Cook's distance after case 13 is deleted). Then MC_2 of the Cook's approach is equal to 1.279, suggesting limited influence. The eigenvector plot given by Cook's approach to local influence (Fig.9.13(a)) shows that cases

Table 9.8: Variables in the Low Birth Weight Data

Variable	Abbreviation
Low Birth Weight (0=Birth Weight $\geq$ 2500g, 1=Birth Weight $<$ 2500g)	LOW
Age of the Mother in Years	AGE
Weight in Pounds at the Last Menstrual Period	LWT
Race (1=White, 2=Black, 3=Other)	RACE
Smoking Status During Pregnancy (1=Yes, 0=No)	SMOKE
History of Premature Labor (0=None, 1=At least one)	PTL
History of Hypertension (1=Yes, 0=No)	HT
Presence of Uterine Irritability (1=Yes, 0=No)	UI
Number of Physician Visits During the First Trimester (0=None, 1=One, 2=Two, etc.)	FTV

13, 148, 172, 184 and possibly 51 are the influential points, although the magnitude of local influence is small. Finally the other panels of Fig.9.13, the eigenvectors of the **MSIF** approach with $\hat{\theta}^2$, LAD and MLE* (MLE with cases 13, 148, 172, 184, 51 deleted) replacing the true parameters also confirm the results provided by Cook's approach to the local influence, although MC_2s are a little different, especially when the true parameters are substituted by MLE* (see table 9.9).

The robust estimates discussed in chapter 6 and 7 are calculated and are listed in table 9.10. Given cases 13, 148, 172, 184, 51 have the potential to be influential, the MLE with those points deleted (MLE*) is also listed as the last entry of table 9.10. Having looked at table 9.10, one may find that all estimates but MLE* are similar, which implies that even though influential points exist in the data, the magnitude of influence is quite small. The final weight of BI estimates that gives 0.75, 0.74, 0.99, 0.87 to cases 13, 148, 172, 184 respectively, also suggests in spite of those four cases departing from the model only a little downweighting is needed to achieve bounded influence. Now, regarding the local influence, MC_2's (see table 9.11), which are almost all between 1 and 2, indicate that the effects of influential points for all those four estimates are small. The only exception is that the

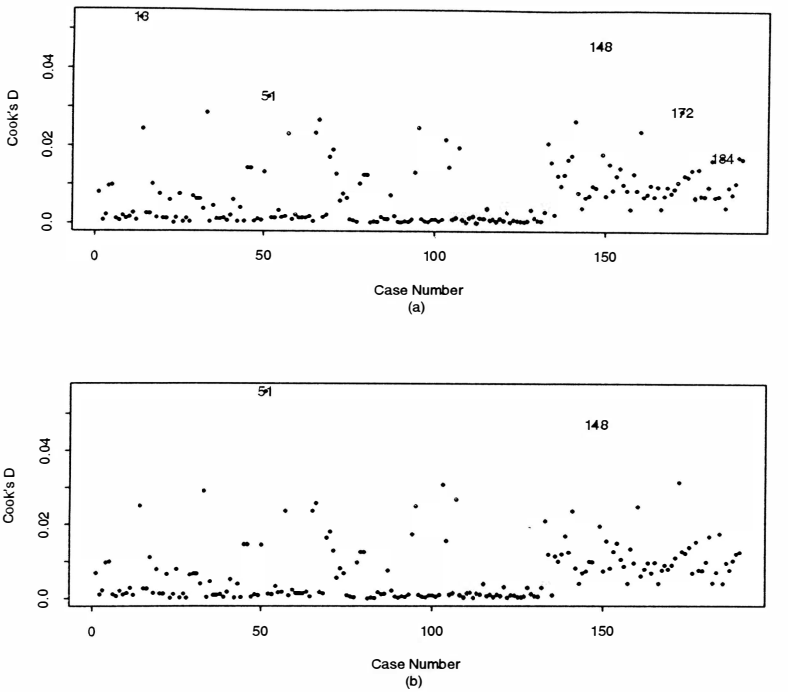

Figure 9.12: Deletion Diagnostics for Low Birth Weight Data, (a) Cook's distance; (b) Cook's distance with case 13 deleted

Table 9.9: Diagnostics for Low Birth Weight Data

Method	Magnitude	Influential Points
Deletion	Maximum Absolute Value	
Cook's Distance	0.0529	13, 148
Local Influence	MC_2	
Cook's Approach	1.279	13, 148, 172, 184 (possibly 51)
Our Approach*	1.364	13, 148, 172, 184, 51
Our Approach**	1.311	13, 148, 172, 184, 51
Our Approach***	3.651	13, 148, 172, 184, 51

* with $\hat{\theta}^2$ replacing the true parameters
** with LAD estimates replacing the true parameters
*** with MLE* replacing the true parameters

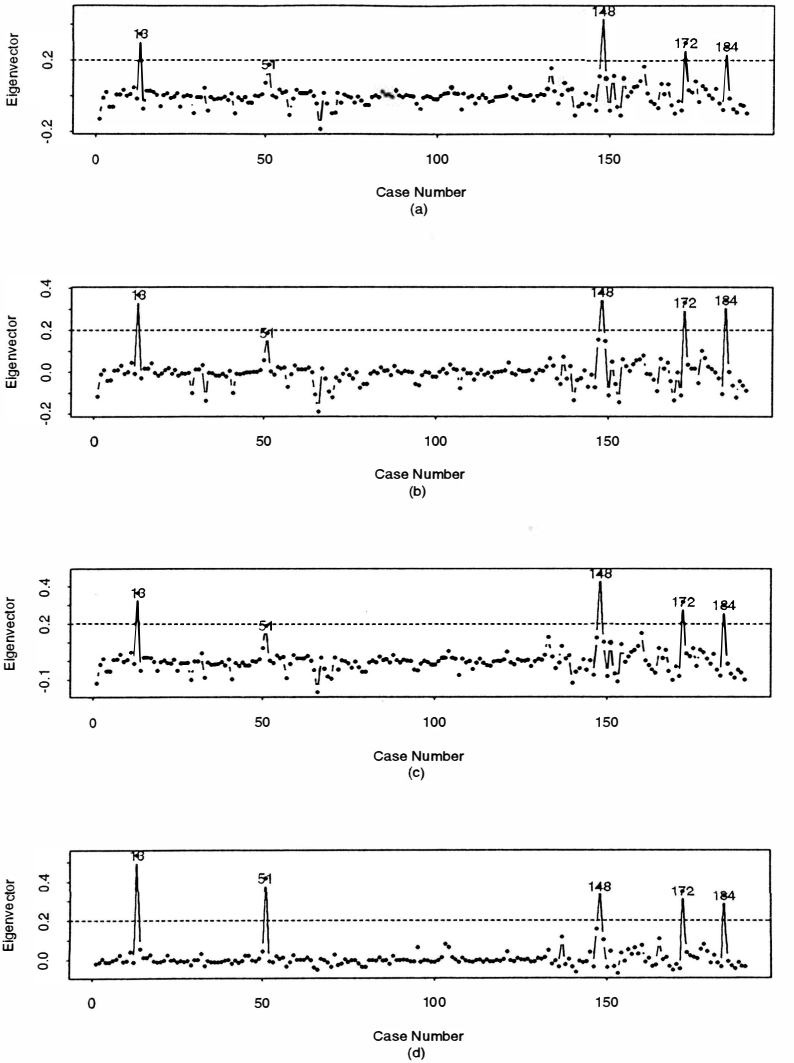

Figure 9.13: Local Influence for Low Birth Weight Data (a) Cook's approach; MSIF approach with true parameters replaced by (b) $\hat{\theta}^2$, (c) LAD estimates and (d) MLE*

Table 9.10: Estimates of Low Birth Weight Data

Estimator	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
MLE	-0.059	-0.016	1.176	0.808	0.811	1.268	1.902	0.735
	(0.98)	(0.01)	(0.52)	(0.44)	(0.40)	(0.46)	(0.71)	(0.46)
Beran	0.281	-0.018	1.147	0.722	0.709	1.316	2.009	0.631
	(1.05)	(0.01)	(0.54)	(0.46)	(0.42)	(0.47)	(0.75)	(0.47)
LAD	0.085	-0.017	1.158	0.755	0.749	1.290	1.929	0.682
	(1.00)	(0.01)	(0.52)	(0.45)	(0.41)	(0.46)	(0.72)	(0.46)
BL	-0.013	-0.017	1.137	0.814	0.807	1.256	1.982	0.699
	(1.000)	(0.007)	(0.522)	(0.443)	(0.404)	(0.461)	(0.735)	(0.463)
BI	-0.026	-0.016	1.149	0.813	0.809	1.257	1.962	0.706
	(0.992)	(0.007)	(0.520)	(0.442)	(0.403)	(0.460)	(0.724)	(0.463)
$\hat{\theta}^1$	-0.447	-0.012	1.017	0.270	0.703	1.812	1.770	1.418
$\hat{\theta}^2$	-0.168	-0.017	1.393	0.694	0.835	1.658	2.127	1.034
MLE*	2.398	-0.035	0.873	0.662	0.469	1.509	3.879	0.519
	(1.28)	(0.01)	(0.57)	(0.46)	(0.44)	(0.50)	(1.11)	(0.49)

* Maximum likelihood estimate with cases 13, 148, 172, 184, 51 deleted

local influence of MLE with MLE* replacing the true parameters, in which MC_2 is 3.487, shows that potentially MLE is influenced by cases 13, 148, 172, 184, 51 (see Fig.9.13(d)). Moreover, in terms of comparing robustness with respect to local influence, it seems hard to give an order of robustness to the Beran, LAD and BI estimators because of the different substitutions for true parameters.

In Fig.9.14-9.16, displays of the eigenvectors of the **MSIF** approach with the three substitutions for the true parameters show that the MLE and BI estimates have been influenced by cases 13, 51, 148, 172, 184 to different degrees (see (a)-(b)). Things are a little different for the Beran and LAD estimates. When $\hat{\theta}^2$ is used, no distinct points show up as influential for both estimates; when LAD estimates are used, cases 13, 51, 148, 172, 184 appear to be influential for both; while when the MLE* is used, only case 51 is influential for Beran's estimates, but cases 13, 51 and possibly some others are influential for the LAD estimates.

Nevertheless, since the magnitudes in table 9.11 are relatively small, *the term of*

Table 9.11: Comparison of Estimates for Low Birth Weight Data

Estimator	MC_2		
	$\hat{\theta}^2$	LAD	MLE*
MLE	1.364	1.311	3.487
Beran	1.666	1.212	1.381
LAD	1.544	1.279	1.951
BI	1.264	1.264	1.551

influential implies to only small changes in parameter estimates and their inference.

So far, we have applied the **MSIF** approach to a few examples in linear regression and logistic regression. Empirical studies, especially for logistic regression, are necessary to examine the effectiveness of this approach and to illustrate the pros and cons of different kinds of robust estimators.

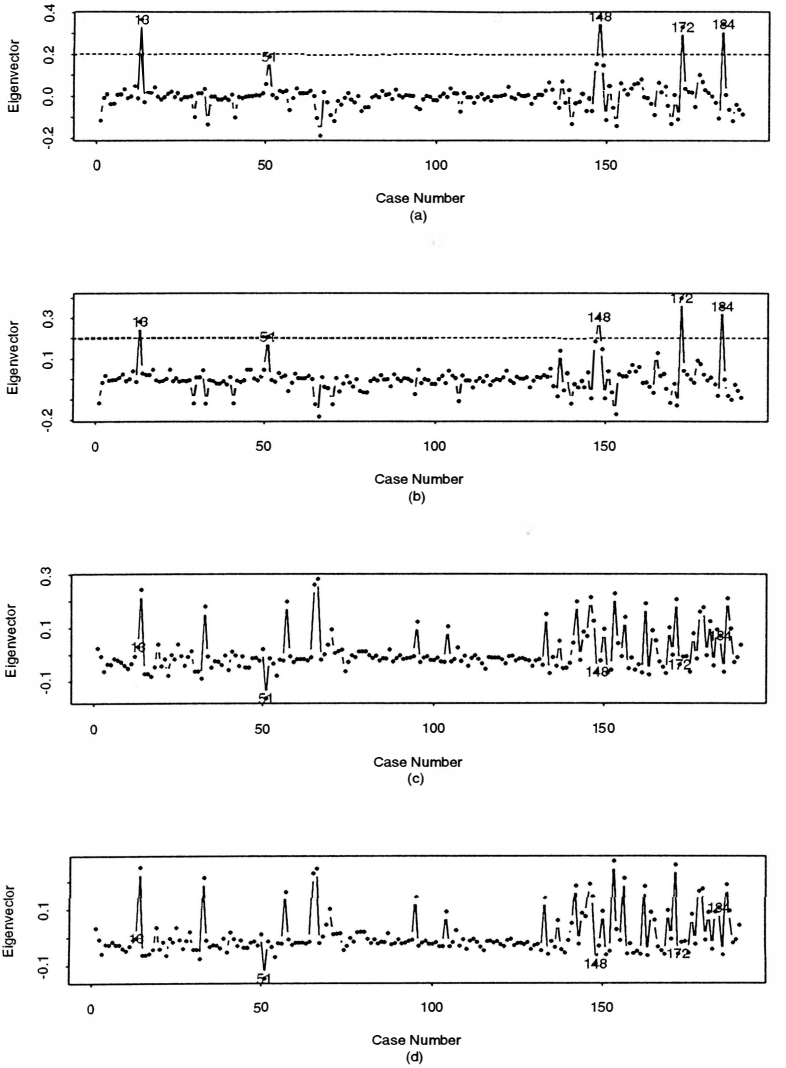

Figure 9.14: Eigenvectors of **MSIF** with $\hat{\theta}^2$ for Low Birth Weight Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

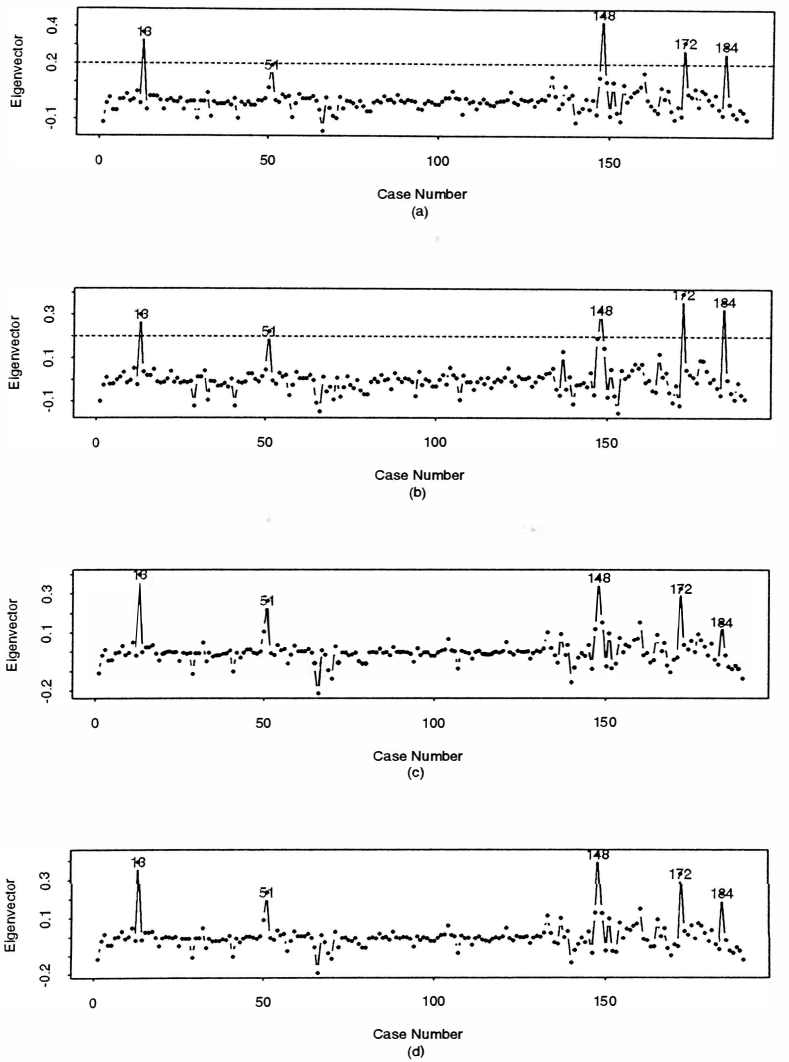

Figure 9.15: Eigenvectors of **MSIF** with LAD estimates for Low Birth Weight Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

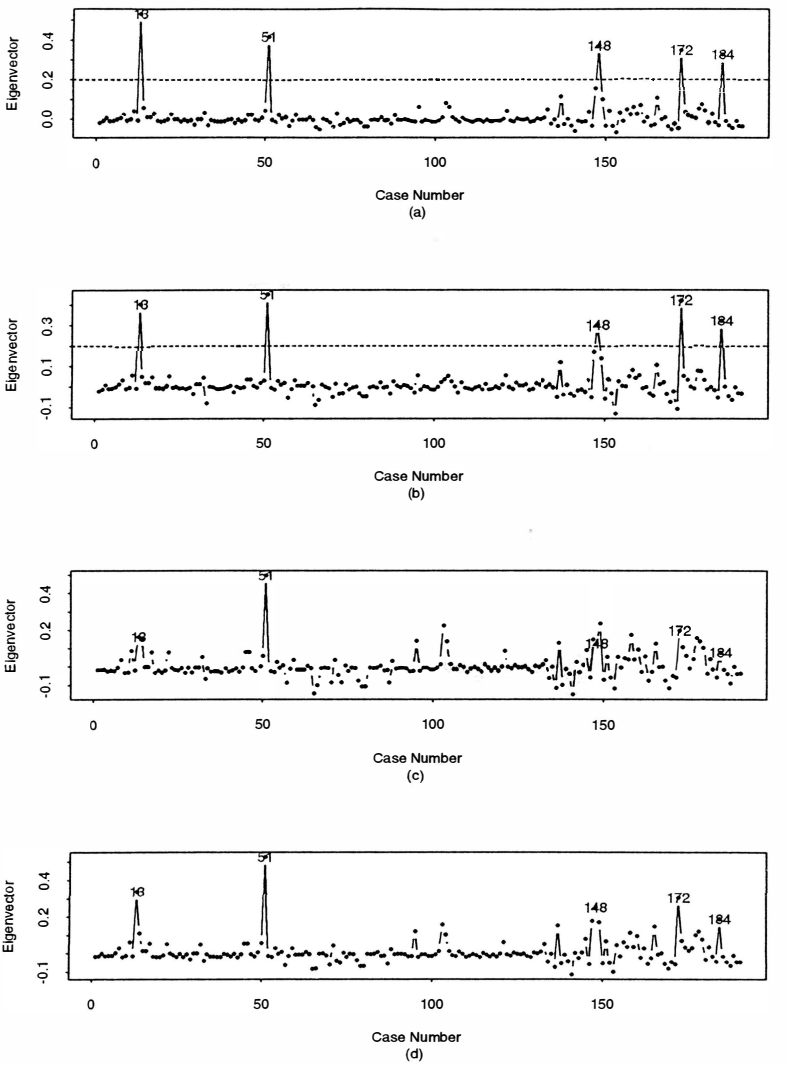

Figure 9.16: Eigenvectors of **MSIF** with MLE* for Low Birth Weight Data, (a) MLE, (b) BI, (c) Beran, (d) LAD

Chapter 10

Conclusions and Future Research

We have developed a measure of local influence based on a standardized influence matrix. As a generalization of Cook's approach to local influence, the standardized influence matrix is an integrated approach to local influence of data for several aspects of statistical inference, for instance, parameter estimates, likelihood and the level of the Wald test statistics. It characterizes the local influence on parameter estimators that can be expressed by functional forms and provides a supplementary measure of the robustness of estimators. It can be used to study the deviation of parameter estimates from true parameters, and will unveil masking effects that plague ordinary diagnostics. Moreover, theoretically it works for all kinds of parametric models and practically at least for some important parametric models.

We have concentrated on its applications to linear regression and logistic regression models. For linear regression, we have examined the least squares estimator, the Huber's robust estimator and the Krasker-Welsch bounded influence estimator. With the least median squares estimator or least trim means squares estimator substituted for the true parameters, we have shown the effectiveness of a diagnostic procedure based on the standardized influence matrix and by this criteria, have found that the Krasker-Welsch estimator is the most robust estimator of the three, while the least squares estimator, not surprisingly, is

the least.

For logistic regression, we have examined the maximum likelihood estimator, Beran's robust estimator, the least absolute deviation estimator and Stefanski's bounded influence estimator. Due to the unique aspects of a binary response, we have shown that it is not feasible to extend the least median squares technique in a logistic regression model. Although a proposed estimator that combines least median squares and smoothing techniques has been proven to be useful in some cases, finding a generally useful robust estimator is a major obstacle to the application of the standardized influence matrix as a diagnostic tool.

There are many related areas of research yet to be done.

1. The concept of bounded local influence needs to be investigated further. Although the relationship between local influence and gross-error sensitivity has been studied, it seems that the concept of bounded local influence is more meaningfully connected with the idea of a breakdown point when the gross-error sensitivity is bounded.

2. The definition of a global robust property or of a breakdown point for logistic regression needs to be revised. Chapter 8 gave the rationale for revision. However, due to the limitation of resources and time we fail to reach a formal definition and give a detailed investigation.

3. Besides applications to linear regression and logistic regression, the standardized influence matrix approach can apply to many other models such as models for polytomous data including ordinal data, models for survival time data and other generalized linear models. Probably each model warrants a Ph.D. dissertation.

4. Further generalizations of the standardized influence matrix and the associated theory are possible. Although the main target of our proposed approach is a possible case

weight perturbation, our methodology can be generalized for other perturbations. For instance,

Small changes in response or explanatory variables.

Since a small change in $\boldsymbol{x}$ can be expressed as

$$F_\epsilon = (1-\epsilon)F + \epsilon[\delta_{x+\Delta x} - \delta_x],$$

$$T(F_\epsilon) \approx T(F) + \epsilon[IF(\boldsymbol{x}+\Delta\boldsymbol{x}, T, F) - IF(\boldsymbol{x}, T, F)].$$

Now if the influence function is continuous and differentiable, and the change that occurs in one variable is small, we have

$$IF(\boldsymbol{x}+\Delta\boldsymbol{x}, T, F) - IF(\boldsymbol{x}, T, F) \approx \Delta\boldsymbol{x}\frac{\partial IF(\boldsymbol{x}, T, F)}{\partial \boldsymbol{x}}.$$

Therefore, if we relax the constraint on the weights to $\pi_i \geq 0, i = 1, \ldots, m$ (see theorem 3.1), and replace the influence function with the derivative of the influence function, the local influence by a small change in a variable will be determined by the generalized standardized influence matrix.

Transposition probability in binary response.

Based on the model given by Copas (1988, see section 2.4.4), we assume that the small transposition probability γ_i occurs in observation $\boldsymbol{x}_i$. Now the influence function of certain estimators by the contamination γ_i at $\boldsymbol{x}_i$ can be developed according to

$$IF(\boldsymbol{x}_i, T, F) = \lim_{\gamma_i \to 0} \frac{T(F_{\gamma_i}) - T(F)}{\gamma_i}.$$

Then the local influence with respect to the transposition probability will be determined by the standardized influence matrix accordingly.

Moreover, we can extend the Copas model into polytomous responses including ordinal variables while assuming the transposition probability happens among adjacent categories of the responses.

With work of this kind based on the standardized influence matrix or a generalization, we can not only identify multiple influential observations but look into relationships among a group of data points and have further insight into the implications of parametric models with less than perfect data.

Bibliography

Bibliography

Agresti, A. (1990) *Categorical Data Analysis* Wiley-Interscience

Atkinson, A. C. (1981) "Two Graphical Display for Outlying and Influential Observations in Regression" *Biometrika.* **68**, 13-20

Atkinson, A. C. (1986) "Masking Unmasked" *Biometrika.* **73**, 533-541

Beckman, R. , Nachtsheim, C. and Cook, R. D. (1987) "New Diagnostics for Mixed Model Analysis of Variance" *Technometrics.* **29**, 413-426

Bedrick, E. J. and Hill, J. R. (1990) "Outlier Tests for Logistic Regression: A Conditional Approach" *Biometrika.* **77**, 815-827

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

Beran, R. (1982) "Robust Estimation in Models for Independent Non-identically Distributed Data" *The Annals of Statistics.* **10**, 415-428

Chambers, J. M. and Hastie, T. J. (1992) *Statistical Models in S* Wadsworth & Brooks

Chatterjee, S. and Hadi, A. S. (1988) *Sensitivity Analysis in Linear Regression* New York: Wiley

Cook, R. D. and Prescott, P. (1981) "Approximate Significance Levels for Detecting Outliers in Linear Regression" *Technometrics.* **23**, 59-64

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

Cook, R. D. (1986) "Assessment of Local Influence" *Journal of the Royal Statistical Society, Ser.B.* **48**, 133-169

Copas, J. B. (1988) "Binary Regression Models for Contaminated Data" *Journal of the Royal Statistical Society, Ser.B.* **50**, 225-265

Cox, D. R. (1989) *The Analysis of Binary Data* London: Methuen

Davies, P. L. (1993) "Aspects of Robust Linear Regression" *The Annals of Statistics.* **21**, 1843-99

Donoho, D. L. and Huber, P. J. (1983) "The Notion of Breakdown Point" in *A Festschrift for Erich L. Lehman* Wadsworth, Belmont, Calif.

Ellenberg, J. H. (1973) "The Joint Distribution of the Standardized Least Squares Residuals From a General Linear Regression" *Journal of the American Statistical Association.* **68**, 941-43

Escobar, L. A. and Meeker, W. Q. (1988) "Using the SAS System to Assess Local Influence in Regression Analysis with Censored Data" *Proceedings of the Thirteenth Annual SAS Users Group International Conference* SAS Institute, Inc., Cary, N.C. 1036-1041

Escobar, L. A. and Meeker, W. Q. (1992) "Assessing Influence in Regression Analysis with Censored Data" *Biometrics.* **48**, 507-528

Finney, D. J. (1947) "The estimation from individual records of the relationship between dose and quantal response" *Biometrika.* **34**, 320-334

Fowlkes, E. B. (1987) "Some diagnostics for binary logistic regression via smoothing" *Biometrika.* **74**, 503-515

Gart, J. J. , Pettigrew, H. M. and Thomas, D. G. (1985) "The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses" *Biometrika.* **72**, 179-190

Gart, J. J. , Pettigrew, H. M. and Thomas, D. G. (1986) "Further Results on the effect of bias, variance estimation and non-normality of the empirical logit on weighted least squares analysis" *Communications in Statistics A, Theory and Methods.* **15**, 755-82

Graham, A. (1987) *Nonnegative Matrices and Applicable Topics in Linear Algebra* Ellis Horwood Limited

Gray, J. B. (1989) "The Internal Norm Approach to Influence Diagnostics" *Communications in Statistics A, Theory and Methods.* **18**, 943-958

Hampel, F. R. , Ronchetti, E. M. and Rousseeuw, P. J. (1986) *Robust Statistics: The Approach Based on Influence Functions* John Wiley & Sons

Hampel, F. R. (1968) "Contributions to the theory of robust estimation" Ph.D. thesis. UC, Berkley

Hampel, F. R. (1971) "A General Qualitative Definition of Robustness" *The Annals of Mathematical Statistics.* **42**, 1887-1896

Hampel, F. R. (1974) "The Influence Curve and its Role in Robust Estimation" *Journal of the American Statistical Association.* **69**, 383-393

Hampel, F. R. (1978) "Optimally Bounding the Gross-error-sensitivity and the Influence of Position in Factor Space" *Proceedings of the ASA Statistical Computing Section*, ASA, Washington, D.C., 59-64

Handschin, E. , Kohlas, J. and Fiechter, A. (1975) "Bad Data Analysis for Power System State Estimation" *IEEE Trans. Power Appar. Sys.* PAS-94. **2**, 329-337

Hastie, T. J. and Tibshirani, R. J. (1992) *Generalized Additive Models* Chapman and Hall

Hawkins, D. M. , Bradu, D. and Kass, G. V. (1984) "Location of Several Outliers in Multiple Regression Data Using Elemental Sets" *Technometrics.* **26**, 197-208

Hettmansperger, T. P. and Sheather, S. J. (1992) "A cautionary note on the method of least median squares" *American Statistician.* **46**, 79-83

Hodges, J. L. (1967) "Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location" *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, 163-186

Holland, P. W. and Welsch, R. E. (1977) "Robust Regression Using Iteratively Reweighted Least-squares." *Communications in Statistics A, Theory and Methods.* **6**, 813-827

Huber, P. J. (1973) "Robust Regression: Asymptotics, Conjectures, and Monte Carlo" *The Annals of Statistics.* **1**, 799-821

Huber, P. J. (1977) "Robust Methods of Estimation of Regression Coefficients" *Math. Operationsforschung Statist. Ser. Statist.* **8**, 41-53

Huber, P. J. (1983) "Minimax Aspects of Bounded-influence Regression" *Journal of the American Statistical Association.* **78**, 60-80

Jennings, D. E. (1986) "Outliers and Residual distributions in logistic regression" *Journal of the American Statistical Association.* **81**, 987-90

Johnson, R. A. and Wichern, D. W. (1988) *Applied Multivariate Statistical Analysis* Prentice Hall, Englewood Cliffs.

Ko, D. and Chang, T. (1993) "Robust M-estimators on Spheres" *Journal of Multivariate Analysis.* **45**, 104-36

Krasker, W. S. and Welsch, R. E. (1982) "Efficient Bounded-Influence Regression Estimation" *Journal of the American Statistical Association.* **77**, 595-604

Leroy, A. and Rousseeuw, P. J. (1984) "PROGRES: A program for robust regression" Technical Report 201, Center for Statistics and O.R., University of Brussels, Belgium

Lund, R. E. (1975) "Tables for an Approximate Test for Outliers in Linear Models" *Technometrics.* **17**, 473-476

Mallows, C. L. (1973) "Influence Functions" National Bureau of Economic Research Conference on Robust Regression in Cambridge, Massachusetts

Mallows, C. L. (1975) "On Some Topics in Robustness" Unpublished Memorandum, Bell Telephone Laboratories, Murray Hill, New Jersey

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Model (Second Edition)* Chapman and Hall

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

Muirhead, R. J. (1982) *Aspects of Multivariate Statistical Theory* New York: Wiley

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

Paul, S. R. and Fung, K. Y. (1991) "A Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear Regression" *Technometrics.* **33**, 339-348

Pettitt, A. N. and Bin Daud, I. (1989) "Case-weighted Measures of Influence for Proportional Hazards Regression" *Applied Statistics.* **38**, 51-67

Pierce, D. A. and Schafer, D. W. (1986) "Residuals in generalized linear models" *Journal of the American Statistical Association.* **81**, 977-986

Pregibon, D. (1981) "Logistic Regression Diagnostics" *The Annals of Statistics.* **9**, 705-724

Prescott, P. (1975) "An Approximation Test for Outliers in Linear Models" *Technometrics.* **17**, 129-132

Prohorov, Y. V. (1956) "Convergence of Random Processes and Limit Theorems in Probability Theory" *Theory of Probability and its Applications.* **1**, 157-214

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection* John Wiley & Sons

Rousseeuw, P. J. and Yohai, V. (1984) "Robust regression by means of S-estimators" In *Robust and Nonlinear Time Series Analysis* Springer, New York.

Rousseeuw, P. J. and Zomeren, B. C. (1990) "Unmasking Multivariate Outliers and Leverage Points" *Journal of the American Statistical Association.* **85**, 633-651

Rousseeuw, P. J. (1983) "Multivariate Estimation With High Breakdown Point" in *Mathematical Statistics and Applications, Vol B*, Reidel, Dordrecht, The Netherlands

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

Rudin, W. (1976) *Principles of Mathematical Analysis* New York: McGraw-Hill

Ruppert, D. and Carroll, R. J. (1980) "Trimmed Least Squares Estimation in the Linear Model" *Journal of the American Statistical Association.* **75**, 828-838

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics* New York: Wiley

Steele, J. M. and Steiger, W. L. (1984) "Algorithms and complexity for least median of squares regression" Technical Report 286, Series 2. Department of Statistics, Princeton University, Princeton, N.J.

Stefanski, L. A. , Carroll, R. J. and Ruppert, D. (1986) "Optimally bounded score functions for generalized linear models with applications to logistic regression" *Biometrika.* **73**, 413-424

Stefanski, L. A. (1983) "Influence and Measurement Error in Logistic Regression" unpublished Ph.D. thesis

Takeuchi, H. (1991) "Detecting Influential Observations by Using a New Expression of Cook's Distance" *Communications in Statistics A, Theory and Methods.* **20**, 261-274

Thomas, W. and Cook, R. D. (1989) "Assessing Influence on Regression Coefficients in Generalized Linear Models" *Biometrika.* **76**, 741-749

Thomas, W. and Cook, R. D. (1990) "Assessing Influence on Predictions in Generalized Linear Models" *Technometrics.* **32**, 59-65

Tietjen, G. L. , Moore, R. H. and Beckman, R. J. (1973) "Testing for a Single Outlier in Simple Linear Regression" *Technometrics.* **15**, 717-721

Tukey, J. W. (1960) "A survey of sample from contaminated distributions" in *Contributions to Probability and Statistics.* Stanford Univ. Press.

Weissfeld, L. A. (1990) "Influence Diagnostics for the Proportional Hazards Models" *Statistics and Probability Letters.* **10**, 411-417

Welsch, R. E. (1979) "Robust and Bounded-influence Regression" *Proceedings of the 42nd Session of the ISI, Manila.* XLVII, 59-68

Williams, D. A. (1987) "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions" *Applied Statistics.* **36**, 181-191

Vita