Diagnostics for Comparing Robust and Least Squares Fits

Joseph W. McKean & Joshua D. Naranjo
Western Michigan University

Simon J. Sheather
University of New South Wales

June 4, 1997

Abstract

Is a simple least squares (LS) fit appropriate for the data at hand? How different would a more robust estimate be from LS? Is a high breakdown estimator necessary, or is a highly efficient robust estimator sufficient? We propose diagnostics which help answer these questions by measuring the difference in fits between least squares and, successively, a highly efficient robust estimate and a bounded influence robust estimate. Our diagnostic \textit{TDBETAS} measures the overall change in parameter estimates among these three fits, while the casewise diagnostic \textit{CFITS} measures change in individual fitted values. We also propose a plot based on \textit{CFITS} which provides an effective graphical summary of underlying data structure.

\textit{Keywords}: Bounded influence; GR-estimates; High breakdown; Linear model; Outlier; Regression diagnostics; and Rank based methods; R-estimates;

1 Introduction

During the past 15 years, a number of methods have been proposed for fitting linear models that achieve various levels of robustness against outlying data points. Each of these proposed estimators have been shown to do well under a certain type of outlier structure, or under some other violations of the ideal conditions for traditional least squares (LS) analysis. Less attention has been devoted to answering the following questions: (i) how different is the robust fit from least squares fit for these data? (ii) if the robust and LS fits are quite different, which is more appropriate? (iii) which data point(s) in particular cause the robust fit to be different from least squares? (iv) which data point(s) will be fitted differently by the robust and LS fits?
In this paper, we propose diagnostics which help answer these questions. The diagnostics that we propose measure differences between least squares and rank-based (R) methods, but the procedure can be readily adapted to other robust methods as well. Key to our diagnostics is the comparison of LS fit with a highly efficient robust estimate and a bounded influence, positive breakdown estimate. For the highly efficient robust estimate we chose the Wilcoxon R-estimate and for the bounded influence, positive breakdown estimate we chose the GR-estimate. The R-estimate is different from LS primarily because the R-estimate is more resistant to y-outliers, and the GR-estimate is different from the R-estimate primarily because the GR estimate is more resistant against x-outliers. Thus, the presence of y-outliers should be detected by a comparison of LS and R estimates (or fitted values), while the presence of x-outliers may be detected by a significant difference in LS or R and GR estimates (or fitted values).

The literature contains considerable work on outlier detection. Many of these are based on outlier deletion. The LS diagnostic DFFITS, for instance, measures the effect of deleting any data point on its own fitted value. For more detailed discussion of LS diagnostics, see Cook and Weisberg (1980, 1989), and Belsley, Kuh and Welsch (1980). McKean, Sheather and Hettmansperger (1990, 1993) showed that some properties of LS residuals hold for R residuals. Among other things, they proposed a rank-based analogue to DFFITS for measuring the effect of deleting a data point on its fitted value. However, even though R estimates are more resistant to y-outliers than LS estimates, both are sensitive to outliers in x-space. Furthermore, single-outlier deletion methods based on LS and R estimates are vulnerable to the masking effect of multiple outliers.

A significant contribution to the detection of multiple outliers was the introduction of high breakdown estimators such as the LMS (Rousseeuw and Leroy, 1987), GM (Simpson, Ruppert and Carroll, 1992 and Coakley and Hettmansperger, 1993), and GR (Naranjo and Hettmansperger, 1994). A standard residual plot using one of these highly robust fits often exposes clusters of outliers. However, these do not measure the degree by which the outliers potentially affect less robust estimates or fitted values. In data analysis, the investigator often wants to know whether a LS approach is appropriate for the data at hand; and if not, which robust estimator is appropriate and which data points need special attention.

It seems clear that analysis of a data set with possible outliers should include a comparison of LS with more robust methods. The proposed diagnostics in this paper do just that, in two
stages. First, we compare the LS fit with the Wilcoxon R-fit, which is robust against y-outliers only. Then we compare the R-fit with the GR-fit, which is in addition robust against x-outliers. The statistic $TDBETAS$ measures the magnitude of the difference in parameter estimates, while the casewise statistic $CFITS_i$, $i = 1\ldots n$ measure the difference between fitted values. Plots of $CFITS$ provide a good picture of the underlying data structure. We chose R- and GR-based methods, but the procedures proposed here can be adapted to other fitting procedures as well (like M and GM estimates).

2 Notation

Consider the linear regression model $y_i = \alpha + x_i'\beta + e_i$, $i = 1\ldots n$ where $x_i'$ is the ith row of the $n \times p$ centered matrix $X$ of explanatory variables. The traditional least squares estimate $\hat{\beta}_{LS}$ minimizes the dispersion

$$D_{LS}(\beta) = \sum_{i=1}^{n}(y_i - \alpha - x_i'\beta)^2.$$ \hspace{1cm} (2.1)

Let $\sigma^2$ be the common error variance. Under Assumptions (A1)-(A3) of the Appendix, $\hat{\beta}_{LS}$ has asymptotic distribution

$$\hat{\beta}_{LS} \sim AN \left( \beta_0, \sigma^2(X'X)^{-1} \right).$$ \hspace{1cm} (2.2)

The regular (Wilcoxon) R-estimate $\hat{\beta}_R$ minimizes the dispersion

$$D_{R}(\beta) = \sum_{i=1}^{n} \left[ R(y_i - x_i'\beta) - \frac{n+1}{2} \right] (y_i - x_i'\beta),$$ \hspace{1cm} (2.3)

where $R(y_j - x_j'\beta)$ is the rank of $y_j - x_j'\beta$ among $y_1 - x_1'\beta, \ldots, y_n - x_n'\beta$. The function (2.3) is a convex function of $\beta$ and Gauss-Newton type algorithms suffice for the minimization; see Kapenga, McKean and Vidmar (1988). Note that (2.3) is invariant with respect to an intercept term, which can be estimated at a later stage as discussed below. Under assumptions (A1)-(A4) in the Appendix, $\hat{\beta}_R$ has asymptotic distribution

$$\hat{\beta}_R \sim AN \left( \beta_0, \tau^2(X'X)^{-1} \right),$$ \hspace{1cm} (2.4)

where $\tau^{-1} = \sqrt{\frac{12}{\pi}} \int f^2(t) dt$, (Jaöckel, 1972). A consistent estimate of $\tau$ is presented in Koul, Sievers and McKean (1987).
The Wilcoxon estimate has relative efficiency .955 with respect to the LS estimate for normal errors and, generally, has much higher efficiency than LS for underlying error distributions with tails heavier than the normal distribution; see Hettmansperger (1984) for general discussions on this efficiency.

Now consider the generalized rank estimate \( \hat{\beta}_{GR} \) defined as the value which minimizes

\[
D_{GR}(\beta) = \sum \sum b_{ij} |(y_i - x'_i \beta) - (y'_j - x'_j \beta)|
\]

where \( \{b_{ij}\} \) are appropriately chosen weights. This dispersion function is also convex and Gauss-Newton type algorithms suffice for the minimization. When \( b_{ij} = 1 \), it can be shown that \( D_{GR}(\beta) = 2D_{R}(\beta) \) so that \( \beta_{GR} = \beta_{R} \), hence \( \hat{\beta}_{GR} \) is a generalized rank estimator.

In the remainder of the paper, \( \hat{\beta}_{GR}(k) \) will refer to the estimator with Mallows-type weighting scheme \( b_{ij} = w_i w_j \),

\[
w_i = \min \left\{ 1, \left[ \frac{c}{(x_i - \mu)^{\top} S^{-1} (x_i - \mu)} \right]^{1/2} \right\},
\]

where \( \mu \) and \( S \) are the minimum volume ellipsoid (MVE) measures of location and scatter (Rousseeuw and van Zomeren, 1990, 1991). For the computations in this paper, we set the cutoff value \( c \) at the 95th percentile of \( \chi^2(p) \). Note that the severity of downweighting increases with \( k \), with \( k = 0 \) corresponding to the Wilcoxon R-estimator. The notation \( \hat{\beta}_{GR} \) with \( k \) suppressed will refer to \( \hat{\beta}_{GR}(k = 2) \), which achieves bounded influence and positive breakdown; see Simpson, Ruppert and Carroll (1992). A breakdown analysis of \( \hat{\beta}_{GR} \) is given in McKeans, Naranjo and Sheather (1996).

Construct the \( n \times n \) matrix of weights \( W = [w_{ij}] \) with off diagonal elements \( w_{ij} = -(1/n)w_i w_j \) and \( i \)th diagonal \( w_{ii} = (1/n) \sum_{h \neq i} w_i w_h \). Under assumptions (A1)-(A8) in the Appendix, the GR-estimate has asymptotic distribution

\[
\hat{\beta}_{GR} \sim AN \left( 0, \tau^2 (X'WX)^{-1} X'W^2X(X'WX)^{-1} \right),
\]

(Naranjo and Hettmansperger, 1994). Note that if \( w_i \equiv 1 \), then \( W = I_n - (1/n)1'1 \), the centering matrix, so that \( X'X = X'WX = X'W^2X \). Hence (2.7) reduces to (2.4), as it should. The GR-estimate is always less efficient than the Wilcoxon estimate. Further for certain designs this loss in efficiency can be substantial; see McKeans, Sheather and Hettmansperger (1994) for discussion.

The LS, R, and GR estimates provide varying levels of sensitivity to outlying data points; the influence function of the LS-estimate is unbounded in both the \( x \)- and \( Y \)-spaces; the influence
function of the R-estimate is bounded in the Y-space but not in the x-space; while the influence function of the GR-estimate is bounded in both spaces. The influence functions of the R- and GR-estimates are derived in Witt, Naranjo and McKeann (1995).

3 Diagnostics that Differentiate between LS and R-Fits

3.1 Difference between Parameter Estimates

Since \( \hat{\beta}_R \) is robust against \( y \) outliers, the statistic \( \| \hat{\beta}_{LS} - \hat{\beta}_R \| \) may be seen as a measure of overall change in parameter estimates due to points with outlying values. In order to be invariant to the measurement scale, this difference needs to be standardized by some measure of standard error. The following theorem states the asymptotic variance of the difference. Its proof, as well as are all other proofs in this paper, is given in the Appendix.

**Theorem 3.1** Under Assumptions (A1)-(A4) of the Appendix, \( \hat{\beta}_{LS} - \hat{\beta}_R \) is asymptotically normal with mean 0 and variance-covariance matrix

\[
\text{Var}(\hat{\beta}_{LS} - \hat{\beta}_R) = \delta^2 (X'X)^{-1},
\]

where \( \delta^2 = \sigma^2 + \tau^2 - 2\kappa \), and \( \kappa = \sqrt{12} \tau E[e(F(e) - 1/2)]. \)

Since \( E[e(F(e) - 1/2)] = \text{Cov}(e, F(e)) \geq 0 \), we have \( \kappa \geq 0 \). By the Cauchy-Schwarz inequality, \( \kappa \leq \tau \sqrt{E(e^2) \cdot E(\sqrt{12}(F(e) - 1/2)^2)} = \tau \sqrt{\sigma^2 \cdot 1} = \tau \sigma \) so that \( \delta^2 \geq (\tau - \sigma)^2 \geq 0. \)

Outliers have an effect on the LS-estimate of the intercept, \( \hat{\alpha}_{LS} \), also. Hence this measurement of overall difference in the R- and LS-estimates needs to based on the estimates of \( \alpha \), also. Traditionally, the intercept \( \hat{\alpha}_R \) is estimated so that the median of R-residuals is 0. This is equivalent to taking \( \hat{\alpha}_R = \text{med}\{y_i - x_i^T \hat{\beta}_R\} \). This, however, raises a problem since

\( \hat{\mu}_{\text{LS}} \) is a consistent estimate of the mean of the errors, \( \mu_e \), while \( \hat{\alpha}_R \) is a consistent estimate of the median of the errors, \( \bar{\mu}_e \). We shall define the difference in these target values to be

\[
\mu_d = \mu_e - \bar{\mu}_e.
\]

One way to avoid a problem here is to assume that the errors have a symmetric distribution, i.e. \( \mu_d = 0 \), but this is unsavory for developing diagnostics for exploratory analysis. Instead, we consider
measures composed of two parts: one part measures the difference in slope parameters and the other part measures differences in the in estimates of intercept.

The version of Theorem 3.1 that includes intercept is the following theorem. Let \( \tau_s = 1/(2f(\theta)) \), where \( f \) is the error density and \( \theta \) is the median of \( f \).

**Theorem 3.2** Under regularity conditions (A1)-(A4) of the Appendix, \( \hat{\beta}_R^* - \hat{\beta}_{LS}^* \) is asymptotically normal with mean vector \( (\mu_d, 0)' \) and variance-covariance matrix

\[
\text{Var} \left( \begin{pmatrix} \hat{\alpha}_{LS} - \hat{\alpha}_R \\ \hat{\beta}_{LS} - \hat{\beta}_R \end{pmatrix} \right) = \begin{pmatrix} (\delta_s^2/n) & 0 \\ 0 & \delta^2(X'X)^{-1} \end{pmatrix} \tag{3.3}
\]

where \( \delta_s^2 = \sigma^2 + \tau_s^2 - 2\tau_s E(e \cdot \text{sgn}(e)) \).

By the Cauchy-Schwarz inequality, \( E(e \text{sgn}(e)) \leq \sigma \) so that \( \delta_s^2 \geq (\tau_s - \sigma)^2 \geq 0 \). Hence, the parameter \( \delta^2 \geq 0 \).

In order to standardize our diagnostics, we need to estimate the parameters involved in (3.3). To estimate the parameter \( E[e(F(e) - 1/2)] \), note that we can write \( D(\hat{\beta}_R) \) as

\[
\frac{1}{n} D(\hat{\beta}_R) = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{n}{n+1} F_n(y_i - x_i\hat{\beta}_R) - \frac{1}{2} \right] (y_i - x_i\hat{\beta}_R), \tag{3.4}
\]

where \( F_n \) is the empirical function of the residuals. It can be shown that (3.4) is a consistent estimate of the parameter \( E[e(F(e) - 1/2)] \); see McKean et al. (1990) for discussion. Applying a degree of freedom correction, we use \( D(\hat{\beta}_R)/(n - p) \) as our estimate of \( E[e(F(e) - 1/2)] \).

\[
E(e \text{sgn}(e)) = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{\alpha}_R - x_i\hat{\beta}_R) \text{sgn}(y_i - \hat{\alpha}_R - x_i\hat{\beta}_R). \tag{3.5}
\]

An estimate of \( \tau \) is given in Koul et al. (1987). An estimate of \( \tau_S \) is given in McKean and Schrader (1984). The parameter \( \sigma \) can be estimated by \( s \), the square-root of mean square error based on the LS-residuals, i.e., \( \hat{\sigma} = \sqrt{\text{MSE}} \). Using these estimates, an estimate of \( \delta \) can be formulated. Other problems in using this estimate of \( \delta \) to standardize the diagnostics proposed below are discussed in Section 3.3.

Let \( \hat{\beta}_{LS}^* = [\hat{\alpha}_{LS}, \hat{\beta}_{LS}]' \) and \( \hat{\beta}_R^* = [\hat{\alpha}_R, \hat{\beta}_R]' \). Define

\[
TDBETAS_D^*(LS, R) = (\hat{\beta}_{LS}^* - \hat{\beta}_R^*)' A_D^{-1}(\hat{\beta}_{LS}^* - \hat{\beta}_R^*) \tag{3.6}
\]
where \( A_D \) is the covariance matrix (3.3). Note that

\[
TDBETAS_D^*(LS, R) = \begin{bmatrix}
\hat{\alpha}_{LS} - \hat{\alpha}_R \\
\hat{\beta}_{LS} - \hat{\beta}_R
\end{bmatrix}
\begin{bmatrix}
\delta_\alpha^2/n & 0 \\
0 & \delta^2 (X'X)^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\alpha}_{LS} - \hat{\alpha}_R \\
\hat{\beta}_{LS} - \hat{\beta}_R
\end{bmatrix}
\]

\[= (n/\delta_\alpha^2)(\hat{\alpha}_{LS} - \hat{\alpha}_R)^2 + (1/\delta^2)(\hat{\beta}_{LS} - \hat{\beta}_R)^2 X'X(\hat{\beta}_{LS} - \hat{\beta}_R) \]

\[= (n/\delta_\alpha^2)(\hat{\alpha}_{LS} - \hat{\alpha}_R)^2 + (1/\delta^2) \| \hat{Y}_{LS} - \hat{Y}_R \|^2 \]

\[= TDINT^*(LS, R) + TDBETAS_D^*(LS, R) \] (3.7)

where \( \hat{Y} \) denotes the vector of fitted values without intercept. The representation above decomposes \( TDBETAS_D^*(LS, R) \) into two components: the intercept part \( TDINT^*(LS, R) \) and the slope part \( TDBETAS_D^*(LS, R) \). The slope part has asymptotic chi-square distribution with \( p - 1 \) degrees of freedom, since it is a quadratic term involving asymptotic normal random vectors. Similarly, the intercept part has asymptotic chi-square distribution with 1 degree of freedom, if the error density is symmetric. If the error density is not symmetric, then \( E(\hat{\alpha}_{LS} - \hat{\alpha}_R) = \mu_d \) (3.2), so that \((n/\delta_\alpha^2)(\hat{\alpha}_{LS} - \hat{\alpha}_R)^2\) is asymptotically noncentral chi-square with 1 degree of freedom and noncentrality parameter \((n\mu^2)/\delta_\alpha^2\).

Thus, under symmetry, \( TDBETAS_D^*(LS, R) \) behaves like a chi-square random variable with \( p \) degrees of freedom. A benchmark for significant difference in fits could be

\[TDBETAS_D^*(LS, R) > \chi^2_{0.05}(p), \] (3.8)

the upper 5th percentile of chi square. If (3.8) declares LS and R fits significantly different, the decomposition (3.7) may pinpoint where the difference lies. We may consider benchmarks for the intercept difference and slope difference separately, as follows:

\[TDINT^*(LS, R) > \chi^2_{0.05}(1) \]

\[TDBETAS_D^*(LS, R) > \chi^2_{0.05}(p - 1). \] (3.9)

If the intercepts are significantly large but not the slopes, then we examine asymmetry of errors. If the slopes themselves are significantly different, then we may declare that LS and R fits are quite different and proceed to casewise analysis (Section 3.2) or other data exploratory analyses (McKean et al., 1996).
3.2 Difference between Fitted Values

If the overall difference between the LS- and R-estimators is large then usually an investigator is interested in the casewise fitted values. In this section we develop diagnostics for evaluating the degree of difference between LS and R fitted values. Once again we will decompose our diagnostic into an intercept part and a slope part. Let \( \hat{Y}^* \) denote the vector of fitted values with intercept, and \( \hat{Y} \) denote the vector of fitted values without intercept. Then

\[
\| \hat{Y}^*_L - \hat{Y}^*_R \|^2 = \| [1 \ X] \begin{bmatrix} \hat{\alpha}_{LS} \\ \hat{\beta}_{LS} \end{bmatrix} - [1 \ X] \begin{bmatrix} \hat{\alpha}_R \\ \hat{\beta}_R \end{bmatrix} \|^2 \\
= \| (\hat{\alpha}_{LS} - \hat{\alpha}_R) 1 + X(\hat{\beta}_{LS} - \hat{\beta}_R) 1 \|^2 \\
= (\hat{\alpha}_{LS} - \hat{\alpha}_R)^2 1^2 + \| Y_L - \hat{Y}_R \|^2
\]

(3.10)

where the last equality follows because \( 1'X = 0 \) implies that the cross product is zero. Equation (3.10) along with (3.7) suggests the following benchmark for the overall difference in fitted values:

\[
\| \hat{Y}^*_L - \hat{Y}^*_R \|^2 > (\delta^2/n)\chi^2_{05}(1) + \delta^2\chi^2_{05}(p - 1).
\]

(3.11)

If the overall difference in fitted values are large, the next logical analysis is casewise exploration: which of the \( n \) data points are fitted quite differently by LS and R? This issue is related to casewise leverage, which loosely measures the influence of a particular data point on the fitted regression plane. Since \( (1/\delta^2) \| \hat{Y}_L - \hat{Y}_R \|^2 \) is asymptotically a chi-square random variable with \( p - 1 \) degrees of freedom, it seems reasonable to expect the \( i \)th component \( (\hat{y}_{LS,i} - \hat{y}_{R,i})^2/\delta^2 \) to behave like a chi-square random variable with \( (p - 1)/n \) degrees of freedom, \( (\Gamma((p - 1)/2n, 2)) \), under approximately uniform leverage. Thus a benchmark for significant difference in the \( i \)th fitted value is

\[
CFITS_{D,i}(LS, R) = (1/\delta^2)(\hat{y}_{LS,i} - \hat{y}_{R,i})^2 > (p - 1)/n + 2\sqrt{2(p - 1)/n},
\]

(3.12)

i.e., two standard deviations above average value.

3.3 The Singularity Problem

The singularity problem, \( \delta = 0 \), cited above becomes a practical problem in the implementation of the diagnostics (3.7) and (3.12). If the LS and R fits are close, then \( \hat{\tau} \) will be close to \( \hat{\sigma} \). This could lead to a practical singularity. However, a more difficult problem concerns estimators for the
parameters $\sigma$, $E[c(F(e) - 1/2)]$, and $E[c \, \text{sgn}(e)]$. The estimators cited above, $\sqrt{MSE}$, ( ??), and ( 3.4), respectively, are all moment estimators and are not robust; see Witt et al. (1995). Estimators such as MAD for $\sigma$ and $\hat{\tau}$ for $E[c(F(e) - 1/2)]$ are consistent only under restrictive distribution assumptions, but, to be of practical use, diagnostics should be valid under a very wide range of distributions. In preliminary simulation investigations these problems were realized in that the diagnostics were frequently too liberal (i.e. standardizing values were too small) and often negative estimates of $\delta_S$ were obtained.

Note, though, that these problems concern only the standardization of the diagnostics. The “numerators” of the diagnostics $TDBETAS_D$ and $CFITS_D$ are intuitive measures of overall and casewise, respectively, difference in fits. What is needed is a robust standardization of them. Such a standardization is discussed next.

### 3.4 Practical Diagnostics

One solution is to standardize the numerator by the standard error of $\hat{\beta}_R^*$ instead of the standard error of the difference. This produces the following statistic:

$$TDBETAS^*_R(\text{LS}, R) = (\hat{\beta}_{LS}^* - \hat{\beta}_R^*)' A_R^{-1}(\hat{\beta}_{LS}^* - \hat{\beta}_R^*)$$

where

$$A_R = \begin{bmatrix} \tau^2_S/n & 0 \\ 0 & \tau^2 (X'X)^{-1} \end{bmatrix}. \quad (3.14)$$

The estimator $\hat{\tau}$ of $\tau$ proposed by Koul et al. (1987) is a quantile type estimator based on the difference of the residuals and it is robust; see Witt et al. (1995) for a derivation of its influence function. It is consistent under the mild regularity conditions (A2-A4) of the Appendix. Likewise the estimator of $\tau_S$ discussed in McKeen and Schrader (1984) is robust.

The diagnostic $TDBETAS^*_R$ also decomposes into separate intercept and slope terms

$$TDBETAS^*_R(\text{LS}, R) = (n/\tau^2_S)(\hat{\alpha}_{LS} - \hat{\alpha}_R)^2 + (1/\tau^2)(\hat{\beta}_{LS} - \hat{\beta}_R)'X'X(\hat{\beta}_{LS} - \hat{\beta}_R)$$

$$= (n/\tau^2_S)(\hat{\alpha}_{LS} - \hat{\alpha}_R)^2 + (1/\tau^2) \| \hat{Y}_{LS} - \hat{Y}_R \|^2. \quad (3.15)$$

$$= TDINT^*(\text{LS}, R) + TDBETAS^*_R(\text{LS}, R)$$

Even if the R and LS fits of the slope parameters are essentially the same, $TDBETAS^*_R(\text{LS}, R)$ can be large because of asymmetry of the errors, i.e. $TDINT^*(\text{LS}, R)$ is large. This may be
worthwhile knowing but in this article we are more concerned with the difference in fits between LS and R. Hence, we will mostly concentrate on the second part, $TDBETAS_R(LS, R)$.

We need to know when is $TDBETAS_R(LS, R)$ large. Since the standardization does not use the covariance matrix of the difference, $TDBETAS_R$ does not have an asymptotic chi-square distribution. McKean et al. (1996) used a similar standardization to assess the magnitude of difference between $\hat{\beta}_R^*$ and $\hat{\beta}_{GR}^*$. Their proposed benchmark was $4(p + 1)^2/n$. We propose the same benchmark here for $TDBETAS_R(LS, R)$. Simulations (see Section 5) show that this benchmark is conservative over normal errors.

We also look at the correspondingly standardized statistic for difference in $i$th fitted value

$$CFITS_{R,i}(LS, R) = \frac{\hat{y}_{R,i} - \hat{y}_{LS,i}}{SE(\hat{y}_{R,i})},$$

(3.16)

where $SE(\hat{y}_{R,i}) = \hat{\tau}[h_i - (1/n)]$, (the expression in brackets is $i$th leverage of the design matrix). As with $TDBETAS_R$, the standardization of $CFITS_{R,i}(LS, R)$ is robust. Note that this standardization is similar to that proposed for the diagnostic $RDFITS$ by McKean, Sheather and Hettmansperger (1990).

Note that (3.16) standardizes by only one fitted value in the numerator (instead of the difference). Belsley, Kuh, and Welsch (1980) used a similar standardization in assessing the difference between $\hat{y}_{LS,i}$ and $\hat{y}_{LS(i)}$, the $i$th deleted fitted value. They suggested a benchmark of $2\sqrt{(p + 1)/n}$, and we propose the same benchmark for $CFITS_{R,i}(LS, R)$. Having said this, we have found it useful in many cases to ignore the benchmarks and simply look for gaps that separate large $CFITS$ from small $CFITS$ (see the examples presented in Section 6).

Simulations, such as those discussed in Section 5, show that standardization at $R$ works much better than standardization at the difference under a wide range of error and $x$-variable distributions.

4 GR Diagnostics

The R-estimate is robust against outlying $y$-values, so a significantly large difference between LS and R fits may mean the existence of $y$-outliers. However, both LS and R estimates are sensitive to points that are outlying in $x$-space, so that comparing LS and R fits may not be very useful in
detecting the existence of high leverage points. In McKeen et al. (1996), diagnostics TDBETAS were proposed that looked at the difference between R and GR estimates.

In this section, we complete the diagnostic tools for comparing fits by developing the theory for comparing LS and GR fits. Thus, we could, for instance, examine a data point \((x_i, y_i)\) and compare the difference in its fitted value for LS, R (robust in \(y\)), and GR (robust in \(y\) and \(x\)).

The following theorem is the LS-GR analogue of (3.3). Let \(\hat{\alpha}_{GR} = \text{med}\{y_i - x'_i \hat{\beta}_{GR}\}\) and let \(\hat{\beta}^*_{GR} = [\hat{\alpha}_{GR}, \hat{\beta}^*_{GR}]'\).

**Theorem 4.1** Under regularity conditions (A2)-(A9) of the Appendix, \(\hat{\beta}_{LS}^* - \hat{\beta}_{GR}^*\) is asymptotically normal with mean vector \((\mu_d, 0)'\) and variance-covariance matrix,

\[
\text{Var} \begin{pmatrix}
\hat{\alpha}_{LS} - \hat{\alpha}_{GR} \\
\hat{\beta}_{LS} - \hat{\beta}_{GR}
\end{pmatrix}
\approx
\begin{bmatrix}
(\delta^2/n) & 0 \\
0 & \tau^2(X'WX)^{-1}(X'W^2X)(X'WX)^{-1} + (\sigma^2 - 2\kappa)(X'X)^{-1}
\end{bmatrix},
\]

where \(\kappa\) is given after (3.1).

Define

\[
TDBETAS^*_D(LS, GR) = (\hat{\beta}^*_{LS} - \hat{\beta}^*_{GR})' A^{-1}_G(\hat{\beta}^*_{LS} - \hat{\beta}^*_{GR})
\]

(4.2)

where \(A_G\) is the covariance matrix in (4.1). Similar to the previous section, (4.2) can be decomposed into the intercept and slope parts. Since the intercept part has asymptotic \(\chi^2(1)\) distribution (under symmetry of errors) and the slope part has asymptotic \(\chi^2(p-1)\) distribution, then a benchmark for (4.2) is

\[
TDBETAS^*_D(LS, GR) > \chi^2_{0.05}(p).
\]

Differences in fitted value may be measured by

\[
CFITS^*_D,i(LS, GR) = \frac{\hat{y}_{LS,i}^* - \hat{y}_{GR,i}^*}{SE(\hat{y}_{LS,i}^* - \hat{y}_{GR,i}^*)},
\]

(4.3)

where \(SE(\hat{y}_{LS,i}^* - \hat{y}_{GR,i}^*) = [1, x'_i] A_G[1, x'_i]'\).

### 4.1 Practical Diagnostics

These GR diagnostics decompose into an intercept part and a slope part. Following the discussion in Section 3, though, we are more concerned with the slope part. Hence, for the remainder of this section we will consider only the slope part.
The singularity problem discussed in Section 3.3 exists here, also, with the estimate of $A_G$ being nearly singular when the LS and GR fits are close. In practice, standardization by either the LS estimate or the GR estimate works better than standardization at the difference. We propose standardization by R estimate instead. This allows for comparability of $TDBETAS$ and $CFITS$ with the statistics of the previous section, and the use of the same benchmarks. Thus, we propose the following criterion for significant differences between LS and GR fits:

$$TDBETAS_R(\text{LS, GR}) = (\hat{\beta}_{LS} - \hat{\beta}_{GR})' \tau^{-2}(X'X)(\hat{\beta}_{LS} - \hat{\beta}_{GR}) > \frac{4(p + 1)^2}{n}$$ (4.4)  

$$|CFITS_{R,i}(\text{LS, GR})| = \frac{|\hat{y}_{LS,i} - \hat{y}_{GR,i}|}{\tau[h_i - (1/n)]} > 2\sqrt{\frac{p + 1}{n}}.$$ (4.5)

### 4.2 Diagnostics for Measuring Differences between R and GR Fits

In the simulation study and examples discussed next, we will make use of the similar diagnostics for comparing the R- and GR-fits that were proposed by McKean et al. (1996). For completeness we restate them here.

Note that both R and GR estimates use the median of their respective residuals as estimates of the intercept parameter. Hence, in there is no reason to decompose the R-GR diagnostics into their slope and intercept parts.

The diagnostic and criteria for the overall difference between the R- and GR-fit is:

$$TDBETAS_R(\text{R, GR}) = (\hat{\beta}_R^* - \hat{\beta}_{GR}^*)' A_R^{-1}(\hat{\beta}_R^* - \hat{\beta}_{GR}^*) > \frac{4(p + 1)^2}{n}.$$ (4.6)

Diagnostics for comparing the $i$th-fitted case between the R- and GR-fits is given by:

$$|CFITS_{R,i}(\text{R, GR})| = \frac{|\hat{y}_{R,i} - \hat{y}_{GR,i}|}{\sqrt{[1, x_i'A_R[1, x_i']]} > 2\sqrt{\frac{p + 1}{n}}.$$ (4.7)

### 5 Simulation Study

Simulations were run to see how the benchmarks performed for the practical diagnostics: $TDBETAS_R(\text{LS, R})$, (3.13); $TDBETAS_R(\text{LS, GR})$, (4.4); and $TDBETAS_R(\text{R, GR})$, (3.13). The benchmark $4(p + 1)^2/n$ was used for all three diagnostics. Data were generated from the model

$$y_i = \alpha + \beta_1 x_i + e_i , \quad i = 1, \ldots, n,$$ (5.1)
Table 5.1: Frequency of times (out of 500) that overall diagnostic declared significant differences.

<table>
<thead>
<tr>
<th>x-Dist.</th>
<th>Y-dist.</th>
<th>( TDBETAS_R(LS, R) )</th>
<th>( TDBETAS_R(LS, GR) )</th>
<th>( TDBETAS_R(R, GR) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unif</td>
<td>N</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Unif</td>
<td>CN</td>
<td>263</td>
<td>262</td>
<td>0</td>
</tr>
<tr>
<td>CN</td>
<td>N</td>
<td>3</td>
<td>395</td>
<td>409</td>
</tr>
<tr>
<td>CN</td>
<td>CN</td>
<td>99</td>
<td>377</td>
<td>379</td>
</tr>
</tbody>
</table>

where \( \alpha = 0 \) and \( \beta_1 = 1 \). We generated the errors from either a standard normal distribution or a contaminated normal. For the error distributions, we chose the standard normal distribution (N) and a contaminated normal distribution (CN), with percentage of contamination 25% and the contaminated variance set at 25. For distributions of the predictors, we chose the uniform distribution on \((0, 1)\) (Unif) and the same contaminated normal distribution that was used for the errors. We set the sample size \( n \) at 20. Hence all together there were \( 2 \times 2 = 4 \) situations. The number of simulations over each situation was set at 500.

Table 5.1 shows the number of times (out of 500) that each statistic showed a significant difference (value greater than benchmark). Only statistics that are standardized at \( R \) are shown in the Table. As discussed in Section 3.3, statistics standardized at the difference too frequently showed negative definite estimated covariance matrices.

The overall diagnostics performed quite well in this study. Over “good” data, Situation 1 (Uniform \( x \)'s and normal errors), the benchmark performed conservatively. Only 4 times out of 500 did they flag the situation. When errors are contaminated and \( x \)'s are uniform (Situation 2), the LS fit is flagged as different from R and GR over half the time. When \( x \)'s are contaminated (high leverage values are present), the GR-fit is flagged as different from the LS- and R-fits in about 80% of the cases. In comparing the 4th situation with the third, note that when the errors are also contaminated the diagnostic \( TDBETAS_R(LS, R) \) flags 20

6 Discussion and Examples

For the examples below, we considered the diagnostics for differences in fits that were evaluated in the simulation study; i.e., \( TDBETAS_R(LS, R) \), (3.13), \( TDBETAS_R(LS, GR) \), (4.4), and
$TDBETAS_R(R, GR)$, (4.6). We also considered the practical, casewise differences diagnostics: $CFITS_{R,i}(LS, R)$, (3.16); $CFITS_{R,i}(LS, GR)$, (4.5); and $CFITS_{R,i}(R, GR)$, (4.7). For illustration we also considered the diagnostic $TDINT_R(LS, R)$ that measures the difference between the LS and R-estimates of the intercept parameter.

A use of these diagnostics would involve obtaining the LS-, R-, and GR-fits of the data set and examining the overall fit diagnostics $TDBETAS_R(LS, R)$ and $TDBETAS_R(R, GR)$. If these are both small then the fits agree and we could use the highly efficient Wilcoxon fit or the LS fit to carry out analyses. This in no way means that the fit of the data is good, but it does give assurance in applying the usual diagnostics based on the LS-fit or the robust diagnostics proposed by McKean et al. (1990) based on the Wilcoxon fit for model criticism. If at least one of the $TDBETAS_R$ diagnostics is large then the corresponding casewise diagnostics $CFITS_{R,i}$ should be examined. A plot that we have found very informative is that of $|CFITS_{R,i}(R, GR)|$ versus $|CFITS_{R,i}(LS, R)|$ overlaid with the benchmark $2\sqrt{(p+1)/n}$, which we present in the examples below. This is a casewise plot whose purpose is too quickly show aberrant cases. These cases are flagged not for deletion but for further investigation. For example, these are the points to “click-on” in a spin or rotational plot of the data. We will label this plot as the $|CFITS_{R}|$-plot.

McKean, Naranjo and Sheather (1996) discussed a procedure for model criticism based on the diagnostics $TDBETAS_R(R, GR)$ and $CFITS_{R,i}(R, GR)$. When $TDBETAS_R(R, GR)$ is large, the cases which have large values of $CFITS_{R,i}(R, GR)$ are set aside and the remaining cases are refitted. This results in a set of “good” data for which the fits agree and a set of “aberrant” data. Note that once a “good” set of data is established, the aberrant cases can be refit one-at-a-time with the good data and the diagnostic $RDFFIT$ can be used to check if the case is influential. This one-at-a-time refitting avoids the masking effect that a cluster of outliers in the $x$-space can have on the Wilcoxon fit. This is illustrated in the second example below.

The third, and final, example below, is a data set exhibiting curvature, which was discussed in McKean, Naranjo and Sheather (1996). As discussed in Cook, Hawkins, and Weisberg (1992) and McKean, Sheather, and Hettmansperger (1993, 1994), and Hettmansperger, McKean and Sheather (1996) high breakdown estimates can have problems in detecting and fitting curvature. Hence, for a given data set large values of $TDBETAS_R(R, GR)$ may indicate curvature problems instead of outliers. In these cases, the diagnostics $CFITS_R(R, GR)$ will tell which cases to click on using a
rotational plot. Often a careful residual analysis using rotational plots along with our diagnostics.

**Example 6.1 Generated Data**

For these examples we consider generated data for situations similar to those in the above simulations. The generating model has $n = 30$ responses, four predictors, and is of the form

$$ y_i = \alpha + \sum_{j=1}^{4} \beta_j x_{ij} + e_i \; ; \; i = 1, \ldots, 30. \quad (6.1) $$

We considered four data sets. For Data Set 1, the $x_{ij}$’s were iid uniform $(0,1)$, the errors were iid $N(0,1)$, and all the regression coefficients were set at 0; hence, this data set is similar to the first situation in the simulation. The generated data appear in Table 6.2.

The $TDBETAS$-diagnostics are displayed in row 1 of Table 6.3. Note that all three diagnostics are far below the benchmark $(4(p + 1)^2/n)$, which has the numerical value 3.33. The value of $TDINT_R(\text{LS}, R)$ is small, also. Hence, all three fits are very similar. The $|CFITS_R|$-plot is shown in Panel A of Figure 6.1. Note that no points exceed the benchmark $(2\sqrt{(p + 1)/n})$ which is .82. The Wilcoxon residual plot and standard diagnostics analysis showed no abnormalities in the fitted model. Thus for Data Set 1, a user would feel comfortable in using the R- or LS-fit.

Data Set 2 used the same data as Data Set 1 but 10 was added to the first three $y_i$’s. Thus we have three outliers in the $Y$-space which is similar to the second situation in the simulation. The $TDBETAS$-diagnostics are displayed in the second row of Table 6.3. Note that $TDBETAS_R(\text{LS}, R)$ exceeds the benchmark but that $TDBETAS_R(\text{R, GR})$ is quite low. Hence the Wilcoxon and GR fits agree but the Wilcoxon and LS fits disagree. Turning to the $|CFITS_R|$-plot as shown in Panel B of Figure 6.1, we see that the casewise diagnostics $|CFITS_{R,i}(\text{LS}, R)|$ exceeds the benchmark for the first three cases, (these are the circled points on the plot). As this plot shows, the outliers sufficiently impaired the LS-fit so that several other cases were poorly misfit by LS also; see Example 6.2 for further discussion.

For Data Set 3, 2 was added to the first three values of $x_{i4}$ and 2 was also added to the corresponding values of $y_i$. Hence, we have 3 contaminated $x$’s and these 3 cases do not follow the model, which is somewhat similar to the third case simulated. The $TDBETAS$-diagnostics are displayed in the third row of Table 6.3. Not surprisingly, the Wilcoxon and LS fits agree but they both disagree with the GR fit. The $|CFITS_R|$-plot, as shown in Panel C of Figure 6.1, easily determines the three points of discrepancy which are circled.
Table 6.2: Observations for Data Set 1

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.072410</td>
<td>0.202244</td>
<td>0.589573</td>
<td>0.805394</td>
<td>0.55212</td>
</tr>
<tr>
<td>0.752125</td>
<td>0.737985</td>
<td>0.287441</td>
<td>0.376231</td>
<td>2.10044</td>
</tr>
<tr>
<td>0.784292</td>
<td>0.906769</td>
<td>0.471071</td>
<td>0.857984</td>
<td>0.06410</td>
</tr>
<tr>
<td>0.006186</td>
<td>0.099673</td>
<td>0.927904</td>
<td>0.025808</td>
<td>-0.63189</td>
</tr>
<tr>
<td>0.696684</td>
<td>0.161842</td>
<td>0.246017</td>
<td>0.381996</td>
<td>0.11023</td>
</tr>
<tr>
<td>0.912110</td>
<td>0.218719</td>
<td>0.885989</td>
<td>0.214536</td>
<td>-1.29266</td>
</tr>
<tr>
<td>0.225229</td>
<td>0.764307</td>
<td>0.390685</td>
<td>0.687817</td>
<td>-0.69871</td>
</tr>
<tr>
<td>0.700249</td>
<td>0.092964</td>
<td>0.312102</td>
<td>0.215019</td>
<td>-0.55932</td>
</tr>
<tr>
<td>0.852142</td>
<td>0.866510</td>
<td>0.632600</td>
<td>0.699470</td>
<td>3.13751</td>
</tr>
<tr>
<td>0.689920</td>
<td>0.366425</td>
<td>0.309782</td>
<td>0.213514</td>
<td>0.12331</td>
</tr>
<tr>
<td>0.265907</td>
<td>0.724180</td>
<td>0.602374</td>
<td>0.416821</td>
<td>1.78939</td>
</tr>
<tr>
<td>0.665918</td>
<td>0.940982</td>
<td>0.592470</td>
<td>0.189538</td>
<td>0.04759</td>
</tr>
<tr>
<td>0.473714</td>
<td>0.635476</td>
<td>0.708251</td>
<td>0.342581</td>
<td>0.19462</td>
</tr>
<tr>
<td>0.796708</td>
<td>0.101960</td>
<td>0.481037</td>
<td>0.251736</td>
<td>0.46283</td>
</tr>
<tr>
<td>0.660012</td>
<td>0.980645</td>
<td>0.533015</td>
<td>0.528750</td>
<td>0.90568</td>
</tr>
<tr>
<td>0.296931</td>
<td>0.980969</td>
<td>0.893360</td>
<td>0.643891</td>
<td>-0.08666</td>
</tr>
<tr>
<td>0.933446</td>
<td>0.178273</td>
<td>0.471665</td>
<td>0.114157</td>
<td>-1.52030</td>
</tr>
<tr>
<td>0.497254</td>
<td>0.152321</td>
<td>0.153870</td>
<td>0.422356</td>
<td>-0.46208</td>
</tr>
<tr>
<td>0.310208</td>
<td>0.677812</td>
<td>0.535185</td>
<td>0.540272</td>
<td>1.07935</td>
</tr>
<tr>
<td>0.935304</td>
<td>0.424143</td>
<td>0.123071</td>
<td>0.201243</td>
<td>-1.40266</td>
</tr>
<tr>
<td>0.976512</td>
<td>0.072957</td>
<td>0.675791</td>
<td>0.705659</td>
<td>0.99251</td>
</tr>
<tr>
<td>0.486728</td>
<td>0.394055</td>
<td>0.871833</td>
<td>0.052314</td>
<td>1.26477</td>
</tr>
<tr>
<td>0.594579</td>
<td>0.933117</td>
<td>0.285833</td>
<td>0.032889</td>
<td>-0.85817</td>
</tr>
<tr>
<td>0.550978</td>
<td>0.409447</td>
<td>0.865447</td>
<td>0.844488</td>
<td>0.48912</td>
</tr>
<tr>
<td>0.857135</td>
<td>0.448007</td>
<td>0.655211</td>
<td>0.166951</td>
<td>-0.57010</td>
</tr>
<tr>
<td>0.424053</td>
<td>0.028971</td>
<td>0.879222</td>
<td>0.603219</td>
<td>-0.69781</td>
</tr>
<tr>
<td>0.207849</td>
<td>0.666022</td>
<td>0.816595</td>
<td>0.226842</td>
<td>-2.08161</td>
</tr>
<tr>
<td>0.132217</td>
<td>0.938948</td>
<td>0.151421</td>
<td>0.905807</td>
<td>0.95338</td>
</tr>
<tr>
<td>0.032062</td>
<td>0.135653</td>
<td>0.018136</td>
<td>0.323158</td>
<td>1.48759</td>
</tr>
<tr>
<td>0.989935</td>
<td>0.247820</td>
<td>0.800386</td>
<td>0.609971</td>
<td>0.70987</td>
</tr>
</tbody>
</table>
Table 6.3: The diagnostics $TDBETAS$ for Data Sets 1-4

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$TDINT_R(LS, R)$</th>
<th>$TDBETAS_R(LS, R)$</th>
<th>$TDBETAS_R(R, GR)$</th>
<th>$TDBETAS_R(LS, GR)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.025</td>
<td>.046</td>
<td>.461</td>
<td>.260</td>
</tr>
<tr>
<td>2</td>
<td>.040</td>
<td>7.89</td>
<td>.074</td>
<td>8.45</td>
</tr>
<tr>
<td>3</td>
<td>.004</td>
<td>.039</td>
<td>3.48</td>
<td>2.89</td>
</tr>
<tr>
<td>4</td>
<td>2.76</td>
<td>4.41</td>
<td>7.34</td>
<td>14.71</td>
</tr>
</tbody>
</table>

Finally for Data Set 4, we contaminated the first three $x_{41}$’s by adding 3 to each of them and adding 2 to their corresponding responses. We also added 15 to $y_{11}$. Hence we have outliers in both the $x$- and $Y$-spaces. As the diagnostics $TDBETAS$ show in the fourth row of Table 6.3 show, all three fits disagree for this data set; even, the difference in the LS and R estimates of the intercept parameter. The $|CFITS_R|$-plot, as shown in Panel D of Figure 6.1, easily determines the three points of contamination in the $x$-space and discovers the point of contamination in the $Y$-space.

**Example 6.2 Case 4 of Example 6.1, continued**

To illustrate the method of model criticism cited above, we will explore further Data Set 4 of the last example. By Panel D of Figure 6.1, the three largest values of $|CFITS_{R,1}(GR, R)|$ are cases 1-3 and there is a very large gap between them and the other few cases that exceed the benchmark. These points, cases 1-3, were “clicked-on” in rotational and scatter plot matrices of the data. The scatter-plot matrix, produced by Cook and Weisberg’s (1994) R-code, is found in Figure 6.2. As the second row, $x_j$’s versus $x_4$, of plots show, these three points form a cluster of outliers in factor space. Hence, by the above method of model criticism, we set these cases aside and fitted the remaining 27 cases. As row 1 of Table 6.4 shows the R- and GR-fits are essentially same on this smaller data set while the LS- and R-fits still differ. The $|CFITS_R|$-plot (contaminated case has been circled), Panel A of Figure 6.3, confirms these diagnostic values.

As Panel A of Figure 6.3 depicts, the R- and LS-fits differ for many cases. We will return to this below, but for now consider the model criticism method of Mckean et al. (1996) applied to the R- and GR-fits. Because the fits agreed on the smaller set of 27 data points, we have confidence in applying the usual diagnostics for robust fits proposed by Mckean et al. (1990) to this smaller
set. The diagnostic $RDFITS$ showed no points of influence while the diagnostic REXT, robust version of the external $t$-diagnostic, showed one large outlier, which was the contaminated $Y$. Next we refit the three outlying $x$-cases added one-at-a-time to the smaller data set. The R-diagnostic $RDFITS$ showed that each of these points were highly influential; their $RDFITS$ values were $-3.14$, $-1.66$, and $-3.74$, respectively, compared with the benchmark of $2\sqrt{(p + 1)/n} = .85$. In contrast when all 30 points of this data set, (Data Set 4), were fit, only one of these points was declared influential by $RDFITS$; their $RDFITS$ values were $-.32$, $-.69$ and $-1.35$, respectively. Hence, the masking effect was avoided.

Now, returning to the difference between the R- and LS-fits, as Figure 6.3 shows, there are many cases where the R- and LS-fits differ. The largest difference is the contaminated $Y$. There is somewhat of a gap between its value for $|CFITS|$ and the next cluster; although, we could also consider as a group the largest 5. This contaminated point can also be seen as the outlier in the top row of plots of the scatter-plot matrix, Figure 6.2. Upon deleting the contaminated $Y$, as the second row of Table 6.4 shows, there is no difference in the resulting three fits. This can also be seen in the $|CFITS_R|$ plot found in Panel B of Figure 6.3. The contrast between the two plots, Panel A and Panel B, is striking. It shows how one outlier can severely foil the LS-fit.

Hence for this data set, the majority of the cases, 4-30, follow a model and there is a cluster of three points, cases 1-3, that are outliers in factor space. There is also one outlier in $Y$-space which is not influential to the R-fit.

### 7 Conclusion

We have proposed diagnostics that help the linear model analyst determine the difference between robust and least squares fits. By exploiting the respective unique sensitivity properties of the least squares, R and GR estimator, the diagnostics $TDBETAS_R(LS, R)$ and $TDBETAS_R(LS, GR)$ not
only declare the presence of outliers, they indicate whether the outliers are in the $x$ or the $y$-space, (see Example 6.1, Table 6.3). These diagnostics measure, respectively, the differences between a highly efficient robust fit and a bounded influence robust fit and between a highly efficient robust fit and a LS fit. If both of these diagnostics are small the analyst can decide whether to use the LS or the highly efficient robust fit; if $TDBETAS_R(R, GR)$ is small and $TDBETAS_R(LS, R)$ is large the analyst knows the outliers reside in the $y$-space; and if $TDBETAS_R(R, GR)$ is large BUT HAVE CURVATURE in TOO aberrant cases Based on these measurements an analyst can decide whether to use the highly efficient robust fit.

For data sets where one or both of these diagnostics, $TDBETAS_R(R, GR)$ and $TDBETAS_R(LS, R)$, is large subsequent analysis using the diagnostics $CFITSR$ help the analyst pinpoint the outliers and points of high influence. We think that the $|CFITSR|$-plot is quite useful in displaying these aberrant cases. The use of $CFITS$ can be quite useful in exploratory or recursive subset-data analysis, where the analyst examines the effects of excluding influential points on the resulting fit. Often, the deletion of a single outlier can dramatically change the level of disagreement between the different estimators, hence uncovering the peculiarities of the data structure (see Example 6.2, Figure 6.3).

As discussed in the examples, these diagnostics used in conjunction with graphic displays such as rotational plots and scatter plot matrices give the analyst powerful tools for exploring data sets. Their use in conjunction with standard diagnostics such as RDFITTS or REXT avoids the masking problem which impairs standard diagnostics.

In this paper, we have intentionally discussed diagnostics as useful tools in the larger problem of model exploration. Often, highly robust methods give the same results as least squares on a subset data with high influence points removed. The issue of whether to use the LS or the robust fit depends on how much the investigator believes the model should fit the high influence points. Resolving these issues should rely, we believe, on subject matter judgement as much as statistical expertise.
8 References


9 Appendix

List of Assumptions:

Let $x_{ik}$ denote the $k$th component of $x_i$. Let $\overline{x}_k = \left(1/\sum w_i \right) \sum w_i x_{ik}$. As $n \to \infty$, assume that

(A1) $F$ has finite variance $\sigma^2$.

(A2) $(\max_i x_{ik}^2) / \sum_i x_{ik}^2 \to 0$, $k = 1, \ldots, p$.

(A3) There exists a $p \times p$ positive definite matrix $\Sigma$ such that $(1/n)X'X \to \Sigma$.

(A4) $F$ has a uniformly continuous, bounded density $f$, such that $f > 0$ a.e. Lebesque, and $f(x) \to 0$ as $x \to \pm \infty$.

(A5) $\max_i |x_{ik}|/\sqrt{n} \to 0$, $k = 1, \ldots, p$.

(A6) \[ \left[\sum_i w_i^2 (x_{ik} - \overline{x}_k)^2 \right] / \left[\sum_i w_i^2 (x_{ik} - \overline{x}_k)^2 \right] \to 0, \quad k = 1, \ldots, p. \]

(A7) \[ \left[\sum_i w_i^2 \sum w_i^2 x_{ik}^2 \right] / \left[\sum_i w_i^2 \sum w_i^2 (x_{ik} - \overline{x}_k)^2 \right] \to 0, \quad k = 1, \ldots, p. \]

(A8) \[ \left[\sum_i w_i^2 \sum w_i^2 x_{ik}^2 \right] / n^2 \text{ is bounded, } k = 1, \ldots, p. \]

(A9) There exist positive definite matrices $C$ and $V$ such that $(1/n)X'WXX \to C$ and $(1/n)X'W^2X \to V$.

Proof of Theorem 3.1

Let $X_1 = [1, X]$ denote the matrix of centered explanatory variables with a column of ones. Recall that $\hat{\beta}_{LS} = (X_1'X_1)^{-1}X_1'Y_1$. Since $1'X = 0$, it can be shown that the vector of slope parameters $\hat{\beta}$ satisfies $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$. Since $Y = \alpha 1 + X\beta + e$, we get the relation

$$\hat{\beta}_{LS} = \beta + (X'X)^{-1}X'e.$$  \hspace{1cm} (9.1)

From McKeon et al. (1990), we have the equivalent relation

$$\hat{\beta}_R = \beta + \sqrt{12\tau} (X'X)^{-1}X'F_c(e) + o_p(n^{-1/2})$$  \hspace{1cm} (9.2)

where $F_c(e) = [F_c(e_1) - 1/2, \ldots, F_c(e_n) - 1/2]'$ is an $n \times 1$ vector of independent random variables. Now,

$$\text{Var}(\hat{\beta}_{LS} - \hat{\beta}_R) = \text{Var}(\hat{\beta}_{LS}) + \text{Var}(\hat{\beta}_R) - 2E(\hat{\beta}_{LS} - \beta)(\hat{\beta}_R - \beta)'$$

$$= \sigma^2(X'X)^{-1} + \sigma^2(X'X)^{-1} - 2\sqrt{12\tau} E[(X'X)^{-1}X'eF_c(e)X(X'X)^{-1}]$$

$$= \delta(X'X)^{-1}$$

22
where \( \delta = \sigma^2 + \tau^2 - 2\sqrt{12\tau E[e_1(F(e_1) - 1/2)].\)

Finally, note that from expressions (9.1) and (9.2) both \( \hat{\beta}_{LS} \) and \( \hat{\beta}_R \) are both functions of the errors \( e_i \). Hence, asymptotic normality follows in the usual way by using (A2) to show that the Lindeberg condition holds.

**Proof of Theorem 3.2**

Recall that \( \hat{\alpha}_{LS} = \hat{Y} = (1/n) \sum_{i=1}^n y_i \). Defining the R-intercept as the median of residuals, it can be shown that

\[
\hat{\alpha}_R = (1/n) \tau_s \sum_{i=1}^n \text{sgn}(y_i) + o_p(n^{-1/2}),
\]

which gives

\[
\text{Var}(\hat{\alpha}_{LS} - \hat{\alpha}_R) = (1/n^2) \sum_{i=1}^n \text{Var}(y_i - \tau_s \text{sgn}(y_i))
\]
\[
= (1/n^2) \sum_{i=1}^n [\text{Var}(y_i) + \tau_s^2 \text{Var}(\text{sgn}(y_i)) - 2\tau_s \text{cov}(y_i, \text{sgn}(y_i))]
\]
\[
= (1/n^2) \sum_{i=1}^n [\sigma^2 + \tau_s^2 - 2\tau_s E(e_i \text{sgn}(e_i))]
\]
\[
= (1/n) [\sigma^2 + \tau_s^2 - 2\tau_s E(e_1 \text{sgn}(e_1))].
\]

Next we need to show that the intercept and slope differences have zero covariance (i.e., that the off diagonal term of \( A_D \) is 0). This follows from the fact that \( 1'X = 0 \). Asymptotic normality follows as in the proof of Theorem 3.1.

**Proof of Theorem 4.2**

Since the intercept \( \hat{\alpha}_{GR} \) is defined as the median of residuals, the linear approximation is the same as \( \hat{\alpha}_R \), so that the asymptotic variance is the same.

From Naranjo et al. (1994), we have the following approximation for the slope parameters:

\[
\hat{\beta}_{GR} = \beta + (\sqrt{3\tau}/n)(X'WX)^{-1}S(\beta) + o_p(n^{-1/2})
\]

where \( S(\beta) = \sum \sum_{i<j} w_iw_j (x_i - x_j) \text{sgn}(y_i - y_j - (x_i - x_j)'\beta) = \sum \sum_{i<j} w_iw_j (x_i - x_j) \text{sgn}(e_i - e_j) \) is a \( p - 1 \times 1 \) random vector. Now,

\[
\text{Var}(\hat{\beta}_{GR} - \hat{\beta}_{LS}) = \text{Var}(\hat{\beta}_{GR}) + \text{Var}(\hat{\beta}_{LS}) - 2E(\hat{\beta}_{GR} - \beta)(\hat{\beta}_{LS} - \beta)'
\]
\[
= \tau^2(X'WX)^{-1}(XW^2X)(X'WX)^{-1}
\]
\[
+ \sigma^2(X'X)^{-1} - 2E[\sqrt{3\tau}/n(X'WX)^{-1}S(\beta)[e'X(X'X)^{-1}].
\]

It can be shown using an elementwise argument that \( E[S(\beta)e'] = 2nE[e_1(F(e_1) - 1/2)]X'W \) and the result follows. Asymptotic normality follows as in the proof of Theorem 3.1.
Figure 6.1: Plot of $|CFITS_{R,i}(R, GR)|$ versus $|CFITS_{R,i}(LS, R)|$ for the data sets of Example 6.1: Data Set 1 (Panel A); Data Set 2 (Panel B); Data Set 3 (Panel C); and Data Set 4 (Panel D)
Figure 6.2: Scatter Plot Matrix of Data Set 4

<table>
<thead>
<tr>
<th></th>
<th>1.00</th>
<th></th>
<th></th>
<th>15.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>x4</td>
<td></td>
<td></td>
<td></td>
<td>-2.08</td>
</tr>
<tr>
<td>x3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Case Deletions

<table>
<thead>
<tr>
<th></th>
<th>0.981</th>
<th>0.928</th>
<th>0.0181</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>x4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.99</th>
<th>0.00519</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>x1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.029</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-2.08</td>
</tr>
</tbody>
</table>
Figure 6.3: The $|CFITS_R|$-Plot for the two cases of Example 6.2: Case 1 (Panel A) and Case 2 (Panel B)