AN EFFICIENT AND HIGH BREAKDOWN PROCEDURE
FOR MODEL CRITICISM

Joseph W. McKeans & Joshua D. Naranjo
Western Michigan University
Kalamazoo, Michigan 49001

Simon J. Sheather
University of New South Wales

Keywords: Diagnostics; GR-Estimates; Influential points; Outliers; R-Estimates; Residual plots; Robust; Wilcoxon-Estimates.

ABSTRACT

Part of a linear model analysis is the examination of the appropriateness of the chosen model. We propose an exploratory model criticism procedure that exposes hidden outliers, clusters of outliers, or underlying curvature by using diagnostics that exploit the differences between an efficient robust fit and a high breakdown fit. Examples illustrate the procedure.

1. INTRODUCTION

In practice a model is often posed in an attempt to describe the relationships between responses variables and predictor variables. It is usually of the form,

\[
\text{responses} = (\text{functions of predictors}) + \text{errors}.
\]

The major assumption is that this indeed is a true model; that is, the errors are random noise. In particular, the errors are independent of the functions of predictors. To test this assumption, observations are obtained and the model is fit by some procedure. This produces estimates of the first term on the right side of the model statement (fits) which in turn results in estimates of the errors (residuals). We then use procedures on the residuals and fits to test the major assumption. These procedures are typically exploratory in nature. They not only focus on how “good” the fit is but on finding outliers and/or clusters of outliers. These are the points that have not been fit well by the model and are often points which lead to interesting discoveries. They may have also had a detrimental influence on the fit itself, thus jeopardizing any conclusions (statistical inference) based on the fit.

In this paper we are concerned with a linear model for one response variable, although the procedures we discuss can easily be generalized to a vector of response variables. The
least squares (LS) procedure offers a rich methodology with which to fit the model and
diagnostic procedures to check the fit and discover outliers; see, for example, Cook and
Weisberg (1982) and Belsley, Kuh and Welsch (1980). It is well known by now that outliers
can severely impair the LS fit and the associated diagnostics. Clusters of outliers can
easily fool one-at-a-time influential LS diagnostic techniques, jeopardizing the discovery of
outliers and points of high influence. Furthermore, computation of multiple case diagnostic
techniques is unfeasible for moderate to large sample sizes and number of predictors unless
the size of the cluster of outliers is known in advance. Newer techniques which have shown
some promise here are those proposed by Hadi and Simonoff (1993).

Robust fitting procedures based on the classes of M-estimates (Huber, 1972) and R-
estimates (Jaeckel, 1972) of regression coefficients were proposed to overcome the sensitiv-
ity of LS fits to outliers. McKean, Sheather and Hettmansperger (1990, 1993) proposed
diagnostic procedures for these fits. They showed that residual analyses for these fits can
proceed analogous to their LS counterparts. These robust fitting procedures and their
associated diagnostics are highly efficient methods relative to the methods based on LS-
estimates. They offer protection to outliers in response space but are sensitive to clusters
of outliers in factor space. High breakdown fitting procedures have been proposed which
are less sensitive to outliers in factor space; see Rousseeuw (1984), Rousseeuw and von
Zomeren (1990), Simpson, Ruppert and Carroll (1992), and Naranjo and Hettmansperger
(1994). These procedures, though, can have problems in detecting and fitting curvature;
see McKean, Sheather and Hettmansperger (1993, 1994) and Cook, Hawkins and Weisberg

McKean, Naranjo and Sheather (1995) recently proposed diagnostics based on both a
highly efficient robust estimator and a high breakdown estimator. These diagnostics can
expose the outliers and/or clusters of outliers which are highly influential to a fit of the
model. They can also expose discrepancies due to curvature in the data. In this paper,
we propose an exploratory procedure based on these diagnostics. After establishing some
notation in Section 2 and briefly reviewing the diagnostics in Section 3, we present our
procedure in Section 4. In Section 5, we use this procedure to explore several data sets.
These data sets exhibit the problems described above. One of the data sets has outliers
only in response space, several have clusters of outliers in factor space, and two of them
contain curvature. In all cases, our procedure is successful in exposing these problems.
Influential outliers and/or clusters of outliers are easily found as well as discrepancies due
to curvature. Residual plots and residual analyses are used throughout so that quality of
fit is also determined. As we present it, it is not an expert system. It allows practitioners
to use subject knowledge in making decisions and it can lead to an enrichment of this
knowledge, also. We think that this procedure is a powerful exploratory tool in model
criticism.

2. NOTATION

Consider the linear model,
\[ Y = \alpha 1 + X\beta + e, \]  \hspace{1cm} (2.1)
where \( Y' = (Y_1, \ldots, Y_n) \) are the observations, \( 1' = (1, \ldots, 1) \), \( X \) is an \( n \times p \) centered,
full rank design matrix, \( \beta' = (\beta_1, \ldots, \beta_p) \) and \( e' = (e_1, \ldots, e_n) \) are iid errors from some
absolutely continuous distribution \( F \) with density \( f \) and median 0.

In this paper we consider R-estimates of the regression coefficients formed under Wilcoxon
scores. These estimates minimize the dispersion function given by
\[ D_R(\beta) = \sum_{i=1}^{n} a(R(Y_i - x'_i \beta))(Y_i - x'_i \beta), \]  \hspace{1cm} (2.2)
where \( x'_i \) is the \( i \)th row of \( X \), \( R(u_i) \) is the rank of \( u_i \) among \( (u_1, \ldots, u_n) \), \( a(i) = \varphi(i/(n+1)) \)
for a non-decreasing function \( \varphi \) defined on the interval \((0, 1)\). In this paper we will only
consider the Wilcoxon score generating function which is given by \( \varphi(u) = \sqrt{12}(u - 1/2) \).
Besides R-estimates, we will sometimes call these estimates the Wilcoxon estimates in this
paper. Dispersion function (2.2) was introduced by Jaeckel (1972) who showed that it
was a nonnegative, continuous and convex function of \( \beta \). We will denote the minimizing
value by \( \hat{\beta}_R \).

Note that the dispersion function is invariant to the intercept. The intercept parameter
can be estimated as a location estimate based on the residuals \( \hat{e}_R = Y - X\hat{\beta}_R \). For this
paper we will estimate \( \alpha \) by the median of the residuals, i.e.,
\[ \hat{\alpha}_R = \text{med} \hat{e}_R. \]  \hspace{1cm} (2.3)
Let \( \hat{\beta}^\star_R = (\hat{\alpha}_R, \hat{\beta}'_R)' \) denote the parameter estimates including the intercept. Under
regularity conditions, Heiler and Willers (1988) and McKean and Hettmansperger (1978)
showed that
\[ \hat{\beta}^\star_R \text{ is asymptotically distributed as } N_{p+1}(\beta^\star, A_R), \]  \hspace{1cm} (2.4)

where

\[ A_R = \text{Cov} \left( \begin{bmatrix} \hat{\alpha}_R \\ \hat{\beta}_R \end{bmatrix} \right) = \begin{bmatrix} \tau^2/n & 0 \\ 0 & \tau^2(X'X)^{-1} \end{bmatrix}, \]

(2.5)

and

\[ \tau = \left( \sqrt{\frac{1}{2n}} \int f^2(x) \, dx \right)^{-1}, \]

\[ \tau_1 = (2f(0))^{-1}. \]

(2.6) (2.7)

We will use the consistent estimate of \( \tau \) proposed by KouI, Sievers and McKeaB (1987). We will use as our estimate of \( \tau_1 \), the estimate based on the standardization length of confidence interval for the intercept as discussed in McKeaB and Schrader (1984). We denote these estimates by \( \hat{\tau} \) and \( \hat{\tau}_1 \), respectively. The influence function of \( \hat{\beta}_R \) is bounded in the \( Y \)-space but is unbounded in the \( X \)-space; see Witt, Naranjo and McKeaB (1995).

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

\[ D_{GR}(\beta) = \sum_{i<j} b_{ij} [ (Y_i - x'_i \beta) - (Y_j - x'_j \beta) ], \]

(2.8)

where \( \{b_{ij}\} \) are appropriately chosen weights; see Naranjo and Hettmansperger (1994). When \( b_{ij} \equiv 1 \), it can be shown that \( D_{GR}(\beta) = 2D_R(\beta) \) so that \( \hat{\beta}_{GR} = \hat{\beta}_R \), 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 \( b_{ij} = w_i w_j \),

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

(2.9)

where \( \mu \) and \( S \) are the minimum volume ellipsoid (MVE) measures of location and scatter (Roussseeuw and van Zomeren, 1990). 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. A breakdown analysis of \( \hat{\beta}_{GR} \) is given in McKeaB, Naranjo and SheaBter (1995).

As with R-estimates, the GR-dispersion function is invariant to the intercept. We will estimate the intercept by

\[ \hat{\alpha}_{GR} = \text{med} \hat{\epsilon}_{GRi}, \]

(2.10)

where \( \hat{\epsilon}_{GR} \) denotes the GR-residuals. Let \( \hat{\beta}_{GR}^* = (\hat{\alpha}_{GR}, \hat{\beta}_{GR}' \)' \). Under regularity conditions,

\[ \hat{\beta}_{GR}^* \] is asymptotically distributed as \( N_{p+1}(\beta^*, A_{GR}) \),

(2.11)
where

\[ A_{GR} = \text{Cov} \left( \begin{pmatrix} \hat{\alpha}_R \\ \hat{\beta}_R \end{pmatrix} \right) = \begin{bmatrix} \frac{\tau_1^2}{n} & 0 \\ 0 & \tau^2 (X'WX)^{-1} X'W^2X(X'WX)^{-1} \end{bmatrix}, \]  \hspace{1cm} (2.12)

where \( \tau \) and \( \tau_1 \) are defined in (2.6) and (2.7).

### 3. Diagnostics for R- and GR-Estimates

The assumption of a linear model is major. Thus, a residual analysis needs to be conducted to check the goodness of fit of the model, to detect outlying points and to detect points of influence. Such a residual analysis has become standard for least squares fitting; see, for example, Cook and Weisberg (1982) and Belsley, Kuh and Welsch (1980), McKea, Sheather and Hettmansperger (1990, 1991, 1993) and Naranjo et al. (1994) developed diagnostic tools for robust fitting. We will briefly review these tools, since we use them in the sequel.

#### 3.1. Residual Plots

A standard check on quality of fit is the residual plot; i.e., the plot of residuals versus fitted values. For least squares fitting, when the fitted model is correct, this plot should look like a random scatter of points. On the other hand, patterns in the plot indicate that the model has been misspecified. In order to see if the properties of robust residuals behave in the same way as the LS residuals, McKea et al. (1990, 1993) and Naranjo et al. (1994) investigated residual plots based on robust estimates. In order to summarize their findings, consider the sequence of misspecified models given by,

\[ Y = \alpha 1 + X \beta + Z \delta_n + e, \] \hspace{1cm} (3.1)

where \( Z \) is a centered \( n \times q \) matrix of full column rank whose columns are independent of the columns of \( X \), and \( \delta_n = n^{-1/2} \theta \) for a \( q \times 1 \) vector \( \theta \). Now suppose we fit Model (2.1) when (3.1) is true. For the R- and GR-fits, first order approximations of the residuals and fitted values are given by:

\[ \hat{Y}_R = X \beta + \tau H_c \varphi(F(e)) + H_c Z \delta_n \] \hspace{1cm} (3.2)

\[ \hat{e}_R = e + \alpha 1 - \tau H_c \varphi(F(e)) + (I - H_c)Z \delta_n \] \hspace{1cm} (3.3)

\[ \hat{Y}_{GR} = X \beta + \frac{\sqrt{3}}{n\tau} X(X'WX)^{-1} S_{GR}(\beta) + K_w Z \delta_n \] \hspace{1cm} (3.4)

\[ \hat{e}_{GR} = e + \alpha 1 - \frac{\sqrt{3}}{n\tau} X(X'WX)^{-1} S_{GR}(\beta) + (I - K_w)Z \delta_n , \] \hspace{1cm} (3.5)
where $H_c$ is the projection matrix onto the column space of $X$, $K_m = X(X'WX)^{-1}X'W$ and $S_{GR}(\beta)$ is defined by,

$$S_{GR}(\beta) = \frac{\sqrt{3}}{n+1} \sum_{i<j} b_ib_j \text{sgn}(e_i - e_j)(x_i - x_j).$$

(3.6)

The corresponding results for the least squares fits are given by

$$\hat{Y}_{LS} = X\beta + H_c e + H_c Z\delta_n$$

(3.7)

$$\hat{e}_{LS} = e + \alpha(1 - H_c e) + (I - H_c) Z\delta_n;$$

(3.8)

see Cook and Weisberg (1982).

McKean et al. (1993,1994) and Naranjo et al. (1994) compared these first order approximations by considering examples and simulation studies. Their results indicated that residual plots based on R-estimates typically behave similarly to plots based on least squares estimates. The GR-residual plots, on the other hand, often show a strong negative association, even when the fitted model was correct; hence, these plots are unreliable tools for detecting misspecified models.

### 3.2. ROBUST DIAGNOSTICS

#### 3.2.1. Studentized Residuals

In LS-fitting, identification of outliers is based on standardized residuals. A standardization that is often used is the standard error of the LS-residuals. These have proved useful in diagnostic procedures, since they correct for both the model and the underlying variance. Typically cases are identified as potential outliers if their corresponding standardized residuals exceed 2 in absolute value.

McKean et al. (1990, 1993) obtained standardized residuals based on robust fits, as follows. Suppose that Model (2.1) is correct. Then a first-order approximation of the R-residual is displayed in expression (3.3) above with $\delta_n = 0$. Based on this the variance-covariance of $\hat{e}_R$ is given by

$$\text{var}(\hat{e}_R) = \sigma^2(I - \kappa_1 J - \kappa_2 H),$$

(3.9)

where $J = n^{-1}11'$, $\kappa_1 = (\tau^2/\sigma^2)((2\delta_1/\tau_1) - 1)$, $\kappa_2 = (\tau^2/\sigma^2)((2\delta_2/\tau) - 1)$, and $\delta_1 = E(|e_i|)$, $\delta_2 = E[e_i \phi(F(e_i))]$. In order to obtain the standardized residuals, we need to estimate these parameters. For $\delta_1$ and $\delta_2$ consider the natural moment estimates which, respectively,
are given by \( \hat{\delta}_1 = (n - p)^{-1} \sum |\bar{e}_{R,i}| \) and \( \hat{\delta}_2 = (n - p)^{-1} D_R(\hat{\beta}_R) \). As an estimate of \( \sigma \), we will use the MAD given by \( \hat{\sigma} = 1.483 \text{Med}_i[|\bar{e}_i - \text{Med}_j \bar{e}_j|] \). It follows that an estimate of \( \text{var}(\bar{e}_{R,i}) \) is

\[
\text{var}(\bar{e}_{R,i}) = \hat{\sigma}^2 (1 - \hat{\kappa}_1 n^{-1} - \hat{\kappa}_2 h_{ii}).
\]

This gives the internal R-studentized residuals as

\[
r_{R,i} = \frac{\bar{e}_{R,i}}{s_{R,i}} \text{ for } i = 1, \ldots, n.
\]

Similarly, Naranjo et al. (1994) showed that the approximate variance-covariance matrix of the GR-residuals is given by

\[
\text{var}(\bar{e}_{GR}) = \sigma^2 I - \kappa_3 J - (\kappa_4 I - \kappa_5 J) K'_w + \tau^2 K_w K'_w,
\]

where \( \kappa_3 = 2 \tau_1 \hat{\delta}_1 - \tau_1^2, \kappa_4 = \sqrt{12 \tau_1 \xi}, \kappa_5 = \sqrt{12 \tau_1 \delta_3}, \delta_3 = E[\text{sgn}\{e_1\}\text{sgn}\{e_1 - e_2\}], \) and \( \xi = E[\text{sgn}\{e_1\} - e_1 \text{sgn}\{e_1 - e_2\}] \). Estimates of the last two parameters are given by the following respective estimators: \( \hat{\delta}_3 = n^{-2} \sum \text{sgn}\hat{e}_i \text{sgn}(\bar{e}_i - \bar{e}_j) \) and \( \hat{\xi} = n^{-2} \sum \text{sgn}(\bar{e}_i - \bar{e}_j) \).

Thus the internal GR-studentized residuals are

\[
r_{GR,i} = \frac{\bar{e}_{GR,i}}{s_{GR,i}} \text{ for } i = 1, \ldots, n.
\]

As with LS-standardized residuals, the standardization of the R- and GR-residuals depend on the location in factor space and the underlying variance of the errors. The standardization for the GR-residuals also depends on the weights. Standardized residuals offer the user a practical scale for identification of potential outliers. Besides the usual residual plots they can be used in other plots involving the residuals, for instance, \( q-q \) plots.

The external \( t \)-residuals are another set of standardized LS-residuals. In this case, the standardization is accomplished through fitting the mean shift outlier model; see Cook and Weisberg (1982). In practice, though, an algebraic equality makes refitting the shift model unnecessary. McKean et al. (1991) proposed an analogue of the external \( t \)-residuals based on robust fits. Unlike the LS case, refitting is necessary for the robust case. Asymptotic linearity can be used to simplify the refitting; see McKean et al. (1991) for details. We will refer to these residuals as the robust external \( t \)-residuals, \( REXT \).

### 3.2.2. Influential Cases

As we noted above, the R-estimates have unbounded influence in the \( \beta \)-space; hence, a useful diagnostic for R-estimates is a measure of the influence a case makes on the fit,
especially, for outlying points in factor space. A simple procedure is the difference between
the fitted value of a point and its predicted value when the case is not in the model. In
terms of LS-fitting, this is the diagnostic DFFIT. For R-estimates, McKeans et al. (1990),
considered the analogue,

\[ \text{RDFFIT}_i = \hat{Y}_{R,i} - \hat{Y}_{R}(i), \]

(3.13)

where \( \hat{Y}_{R}(i) \) denotes the R-predicted value of \( Y_i \) when the \( i \)th case is deleted from the
model. To be useful RDFFIT needs to be standardized. The asymptotic variance of the
R-fitted value is

\[ \text{Var}(\hat{Y}_{R,i}) = n^{-1} \tau_1^2 + h_{ci}\tau^2. \]

(3.14)
The scale parameters in this expression can be estimated from the original model or the
delete \( i \) model. Following Belsley et al. (1980) for the analogous LS-diagnostic, we shall
base the estimates on the delete \( i \) model. This leads to the diagnostic

\[ \text{RDFFIT}_i = \frac{\text{RDFFIT}_i}{\sqrt{n^{-1} \tau_1^2 + h_{ci}\tau^2}}, \]

(3.15)

which was proposed by McKeans et al. (1990). As with the external \( t \)-statistics discussed
above, the LS-version of DFFITS can be obtained by an algebraic identity but this is not
true for the R-version; hence, refitting is necessary for it.

3.3. DIAGNOSTICS TO DETECT DIFFERENCES IN ROBUST FITS

Since \( \hat{\beta}_{GR}(k = 0) \equiv \hat{\beta}_R \), the statistic \( \| \hat{\beta}_R - \hat{\beta}_{GR}(k) \| \) where \( k > 0 \) is a measure of
overall change in parameter estimates due to the downweighting of high leverage points. In
order to be useful, this difference needs to be assessed relative to some scale. It would seem
that the asymptotic variance-covariance of the difference \( \| \hat{\beta}_R - \hat{\beta}_{GR}(k) \| \) is a suitable
choice, but, as shown in McKeans, Naranjo and Sheather (1995), it is a singular matrix
due to the inclusion of the estimates of the intercept parameter. Even if the intercept
is omitted, this scaling based on the difference is not sufficiently stable for cases whose
weights are close to one. A solution is to standardize \( \| \hat{\beta}_R - \hat{\beta}_{GR}(k) \| \) by the standard
error of \( \hat{\beta}_R \) instead of the standard error of the numerator. This produces the following
statistic, which measures the total difference in the fits of \( \hat{\beta}_R \) and \( \hat{\beta}_{GR} \),

\[ TDBETAS_R = (\hat{\beta}_R - \hat{\beta}_{GR})' A^{-1}_R (\hat{\beta}_R - \hat{\beta}_{GR}). \]

(3.16)

where \( A^{-1}_R \) is the asymptotic variance-covariance of \( \hat{\beta}_R \) given by expression (2.5). As
discussed in McKeans et al. (1995), an appropriate benchmark is to flag \( TDBETAS_R \) as
large whenever

\[ TDBETAS_R > \frac{4(p + 1)^2}{n}. \]  

(3.17)

A similar standardization can be based on the \( A_{GR} \); however, standardization based on \( A_R \) behaved somewhat better in simulation studies; see McKeen et al. (1995).

The diagnostic \( TDBETAS_R \) measures an overall difference between the R- and the GR-fits. Upon obtaining a large value of it, a user would be inclined to investigate individual cases; i.e., differences in individual fitted values. Let \( \hat{Y}_{R,i} = \alpha_R + \mathbf{x}_i^T \hat{\beta}_R \) and \( \hat{Y}_{GR,i} = \alpha_{GR} + \mathbf{x}_i^T \beta_{GR} \) denote the respective fitted values for the \( i \)th case. A suitable scale to compare these differences against are fitted value scales, similar to the regular diagnostic \( DFFITS \). As stated above, the R-fitted scale performed somewhat better than the GR-fitted scale in a simulation study. This gives

\[ CFITS_{R,i} = \frac{\hat{Y}_{R,i} - \hat{Y}_{GR,i}}{\text{SE}(\hat{Y}_{R,i})} \]  

(3.18)

where \( \text{SE}(\hat{Y}_{R,i}) = \{|1, \mathbf{x}_i^T|A_R|1, \mathbf{x}_i^T|\}^{1/2} \). This type of standardization is similar to \( DFFITS \) and \( RDIFFITS \), and we propose the same benchmark \( 2\sqrt{(p + 1)/n} \). We should note here that the objective of the diagnostic \( CFITS \) is not outlier deletion. Rather the intent is to identify the critical few data points on which the two fits differ greatly for closer study, because these critical few points often largely determine the outcome of the analysis or the direction that further analyses should take. This closer study may involve subject matter expertise, or rotating brushplots, or data-collection-site investigation for accuracy of measurements, among other things (see Examples 2 and 3 of Section 5). In this regard, the proposed benchmark of \( 2\sqrt{(p + 1)/n} \) is meant as a heuristic aid, not a boundary to some formal critical region. The simulation study in McKeen et al. (1995) showed that the benchmark is in fact quite conservative when no outliers are in design space. On the other hand, when large outliers in the \( \mathbf{x} \) space are present the \( CFITS \) value of influential observations tend to be large. In light of this, along with these benchmarks, we have found it useful in many examples to simply look for gaps in size that separate large \( CFITS \) from small \( CFITS \) values; see the examples of Section 5.

### 4. EXPLORATORY MODEL CRITICISM

We believe that the diagnostics proposed in the last section, \( TDBETAS \) and \( CFITS \), are powerful tools for model criticism. The following procedure describes one way that we have made use of these diagnostics in analyzing data sets. This is not an expert system. As
with other diagnostic tools, we strongly believe that procedure described below be used in conjunction with knowledge of the subject matter, graphical methods and common sense.

1. Using R- and GR-estimates, make initial fits on all the data. Obtain the diagnostic $TDBETAS$.

2. If $TDBETAS$ is less than its benchmark, $(3.17)$, this gives evidence that there are no serious outliers in the $x$-space. This in no way says that the model is adequate. But it does give some confidence in using the highly efficient fit to check the adequacy of the fit of the model. This includes the use of residual plots, $q-q$ plots, and standard robust diagnostics, as discussed in Section 3. Such a check may very well lead to the consideration of other models. For example, there may be a curvature pattern in the residual plot or an indication of heteroscedasticity in the residual plot.

3. If $TDBETAS$ exceeds its benchmark then there is evidence that R- and GR-fits substantially differ. The casewise diagnostics, $CFITS$, are next examined to determine what cases were the main contributors to this overall difference. We have found it best to consider the ordered $|CFITS_i|$ and look at the larger values. For instance, gaps between groups of values are worthy of inspection because a cluster of outliers in the $x$-space will often have similar $CFITS_i$ values. We then place the cases with largest values into a group labeled Group B and place all other cases in Group A.

The use of dynamic plots such as brush plots cannot be over emphasized, here. In a brush plot, the cases in Group B are the ones to click on. Often by rotating and changing the axes for the various predictor variables, clusters of points are brought to one’s attention; see Cook and Weisberg (1994). This may be the most important contribution of these diagnostics, leading to an investigation of the subject matter for the rationale behind such clusters.

4. If the preceeding analysis or subject matter indicate that the discrepancies are due to curvature then we may not want to proceed with the next step which involves refitting the model using only the points in Group B. In these circumstances, we may feel comfortable with the highly efficient robust fit. The last example in Section 5 serves as an illustration of this.

5. If we do decide to refit the cases in Group A, using both R- and GR-estimates, we proceed as follows:
(a) If upon refitting, $TDBETAS$ is still large we proceed through Step 3 once more. This will lead to an enlargement of Group B.

(b) If $TDBETAS$ is small or Step (5-a) was performed, the standard R-diagnostics are run on the cases that were set aside in Group B. This is performed by adding these points one at a time to Group A and then obtaining the R-diagnostics. Hence, the masking effect that a cluster of outliers in the $x$-space has on the R-estimate is avoided. The diagnostics to use here are $RDFFITS_{1\alpha}$ (3.15), and $REXT_{1\alpha}$ (see the discussion in Section 3.2.1), to see if an excluded case follows the same model as Group A does.

5. EXAMPLES

We offer several examples to illustrate the diagnostics procedure outlined in the last section. The first example contains outliers in the $Y$-space only. This is followed by three examples with outliers in the $x$-space while the last two examples concern data sets which contain curvature. Our diagnostics presented in Section 3 easily find the sources of difficulty with each of these data sets.

**Example 5.1 Telephone Data**

The response for this data set is the number of telephone calls in Belgium while the sole predictor is the year (1950-1973). During the years 1964-1969 another type of recording system was used; see page 26 of Rousseeuw and Leroy (1987) for a discussion of this data set. The diagnostic $TDBETAS = .0014$ for this data set. This implies the R-, (Wilcoxon), and GR-fits are essentially the same. The scatterplot overlaid with the LS- and Wilcoxon-fits is shown in the left panel of Figure 5.1.

The scatterplot shows that the outliers for this data set are in the $Y$-space and both robust estimates are resistant to such outliers. The LS-fit is severely impaired by the outliers. Note that the Wilcoxon Studentized residuals flagged all of the outlying points. Although not shown, only one of the LS Studentized residuals surpassed the benchmark.

**Example 5.2 Stars Data**

This data set consists of the measurements of the effective temperature at the surface (the predictor) of 47 stars and the light intensity (the response) of the star. The data can be found on page 27 of Rousseeuw and Leroy (1987). The logarithmic transformation was
applied to both the predictor and the response. The following is a brief summary of our procedure applied to this data set:

On the original data \( TDBETAS = 101.5 \) which far exceeds the benchmark. The four largest \( |CFITS| \) had values (to two places) 9.7. The next largest value was 8.7. We placed the cases with the 4 largest values in Group B while the remaining cases were placed into Group A. Upon fitting Group B, the value for \( TDBETAS \) was 2.1 which, though, is much smaller than its value on the original data still exceeded the benchmark of .37. The two largest \( |CFITS| \) values were at 1.3 while the next largest was at 1.0. We added the two largest cases to Group B. This left 41 cases in Group A. Fitting these cases resulted in \( TDBETAS \) at 0.0 (to 5 places); hence, there is strong agreement between the Wilcoxon- and GR-fits based on these 41 cases. Based on the fit with the 41 cases, the one-at-a-time diagnostic \( RDFFITS \) exceeded its benchmark for each of the 6 cases in Group B, indicating that these six points do not fit the same model as the other 41 cases.

With just one predictor, the brush plot is the scatter plot of the data given by the left panel of Figure 5.2. The 4 largest \( |CFITS| \) values correspond to the cluster of 4 points in the upper left corner of the plot. The next group of two largest are the points between the group of 4 and the main group. Subject matter is quite helpful here. The group of 4
Figure 5.2. Left Panel: Wilcoxon (bold line) and GR (dashed line) fits of the Stars Data.  
Right Panel: CFITS versus Case.

are red giant stars while the large group of 41 stars are main sequence stars. The group of 2 stars are transition stars between the two groups. Thus our exploratory diagnostics have been successful in verifying the subject matter. The right panel of Figure 5.2, give the R- and GR-fits and the casewise CFITS; for the fits of all 47 points.

Example 5.3 Hawkins Data

This is an artificial data set developed by Hawkins, Bradu and Kass (1984). It consists of 75 data points and three predictors. It contains two groups of outliers in factor space. The first group, cases 1-10, do not follow the model of cases 15-75 while the second group, cases 11-14, do follow this model.

The diagnostics TDBETAS has the value 1238, indicating a strong disagreement in R- and GR-fits. The largest values of |CFITS|; are for the cases 1-14. These values range from 13.8 to 33.3. The next largest value was 2.65. This information is summarized in the left panel of Figure 5.3. Hence these diagnostics had no trouble in finding the two groups of outliers. The right panel of Figure 5.3 shows the casewise standardized residual from the GR-fit. Note that if we used the information solely from the GR-fit, then only the one group, Cases 1-10, would have been found. Even though the group of cases 11-14 supposedly follow the model, we think that the knowledge of these extreme points in factor space would be useful to a user.
Figure 5.3. Left Panel: CFITS versus Case for the Hawkins Data.
Right Panel: GR-Studentized residuals versus Case.

Table 5.1: Standard Diagnostics on the Outliers Based on GROUP A Data

<table>
<thead>
<tr>
<th>Case</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDFFITS</td>
<td>17.2</td>
<td>17.9</td>
<td>18.2</td>
<td>17.4</td>
<td>18.1</td>
<td>17.5</td>
<td>18.5</td>
<td>18.8</td>
<td>17.5</td>
<td>17.9</td>
<td>3.8</td>
<td>3.8</td>
<td>4.9</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Figure 5.4 depicts a brush plot of the data consisting of predictors 2 and 3 and the response variable. This plot was obtained with the R-code software discussed in Cook and Weisberg (1994). It shows two distinct groups of outliers: Cases 1-10 far removed from the bulk of the data and Cases 11-14 not quite as away from the bulk of the data. The second group appears to conform more with the bulk of the data.

For the second run on this data set, we put all 14 outliers into Group B and obtain R- and GR-fits on the remaining data. For this run, TDBETAS = .61 which is below the benchmark of 1.04; hence, the R- and GR-estimates agree on the bulk (data without Cases 11-14) of the data. Table shows the diagnostic RDFFITS, for each of the outlying cases when they are fitted, one-at-a-time, with the bulk of the data. All the values are high, but note the distinct gap between Cases 1-10 and Cases 11-14.

**Example 5.4 Wood Data**

This is a real data set (Draper and Smith, 1966) which was modified by Rousseeuw (1984) to contain four outliers. It contains 5 predictors and 20 observations.
The diagnostic $TDBETAS$ has the value 280 for this data set, implying a major difference in the R- and GR-fits. Cases 4, 6, 8 and 19 have $|CFITS|$ values of 14.7, 15.7, 15.3 and 16.2, respectively. The brush plot, Figure 5.5, shows that these 4 cases form a cluster of points in factor space. These are the 4 cases which were modified by Rousseeuw. The next largest value of $|CFITS|$ was 4.80 for case 11. Hence, we let Group B consist of the 4 largest $|CFITS|$ values. Our second fit was based on Group A. For the second fit, $TDBETAS$ had the value 2.74 which is less than the benchmark value of 9. Thus the procedure has isolated the points which caused the discrepancy. In practice, hopefully subject knowledge would provide rationale for why these 4 cases are outliers.

Our final examples are concerned with data sets containing curvature. Cook, Hawkins and Weisberg (1992) and McKeen, Sheather and Hettmanperger (1993, 1994) showed that high breakdown estimates can have problems with fitting and detecting curvature. The first example involves one predictor. Based on the scatterplot a quadratic relationship is obvious. Here the high breakdown estimate offers a poor fit. The second example has two predictors. Without a simple scatterplot, one might proceed with a first-order model, as we do.
Example 5.5 Synthetic-Rubber Data

This data set consists of 37 observations from a synthetic-rubber process taken from Mason, Gunst and Hess (1989, p. 488). The response variable is the weight (in percent) of a solvent while the independent variable is the corresponding production rate of the rubber process. For computational convenience we have divided the predictor variable by 100 and then centered it by subtraction of the mean. Robust fitting of this data is discussed in McKeen, Sheather and Hettmansperger (1994). As discussed there, a quadratic model in production rate seems to be appropriate; see the scatterplot, left panel of Figure 5.6, also.

Based on the plotted fits, the GR fit misread the curvature in the plot. Not surprisingly, the value of TDBETAS at 1070 is quite large; hence, this diagnostic verifies the real difference in the R- and GR-fits. As the right panel of Figure 5.6 shows, the diagnostic |CFITS| clearly determines the several cases where the major disagreement between the fitted values occur. The top 6 |CFITS| values ranged from 28 to 32. The next largest value was 20.1. We placed these top 6 cases into Group A. For the second run TDBETAS has a much smaller value of 40.6, but it is still considerably higher than the benchmark of 1.2. Thus the R- and GR-estimates do not agree on the smaller data set.

Since there is only one predictor for this data, the scatterplot, Figure 5.6, serves as
our brush plot. Of course upon viewing this plot it is clear that the GR-fit misread the curvature in the data. On the other hand, the Wilcoxon estimates fit the data well. As discussed in McKean et al. (1993, 1994) high breakdown estimates can have problems fitting curvature. This data set is such an example. It has the same effect on other high breakdown robust estimates; see McKean et al. (1994).

**Example 5.6 Response Type Data**

This is an artificially generated data set. For two predictor variables, say, $x_1$ and $x_2$, we chose a response surface design. Six points were chosen within the circle of radius 1, five points were chosen in the ring with radii 1 and 2, and four points were chosen in the ring with radii 2 and 3. There were two replicates at each of these points; hence, 30 points altogether. As a model we chose

$$y = -9x_1 - 9x_2 + 3x_1^2 + 3x_2^2 - 2x_1x_2 + e,$$

(5.1)

where the errors are simulated iid $N(0,100)$. The $E(Y)$ has a minimum at the point $(2.25,2.25)$.

Suppose we did not know the true model. In this case as a first model, we would fit the first-order model with $x_1$ and $x_2$ as the predictors. Based on the R- and GR-fits
of the linear model, $TDBETAS$ had the value .22 while the benchmark is 1.2. Hence, the R- and GR-fits are similar. The residual plots for the Wilcoxon and GR-fits of the first-order model are found in Figure 5.7. In both plots curvature is indicated. Although not shown, this was true of the LS residual plot also. Hence we would proceed to fit at least a quadratic (second-order) model.

The residual plots of the full second-order model are presented in Figure 5.7, also. The random scatter in the Wilcoxon plot are an improvement over its respective plot for the first-order model. It indicates a better fit with the second-order model. On the other hand, the GR-residual plot is poor. It shows a negative linear pattern which is similar to the patterns discussed in McKean et al. (1993 and 1994) for high breakdown fits in the presence of curvature. The LS plot was similar to the Wilcoxon plot. Both LS and R-fitted response surfaces indicate a minimum value while the GR-fit indicates a saddle point; hence the GR estimate has misread the curvature in this data set.

In light of the residual plots, it is unsurprising that the diagnostic $TDBETAS$ for the quadratic model is high, 1140. For this model, though, the series of residual plots show that the Wilcoxon-fit results in a much better fit to the data than the GR-fit.

6. CONCLUSION

Exploiting the differences between an efficient robust fit and a high breakdown fit, we have proposed an exploratory model criticism procedure that exposes hidden outliers, clusters of outliers, or underlying curvature. The examples in Section 5 illustrate how the procedure works. Plots, diagnostics, benchmarks, and common sense help to determine significant differences in fits and significantly influential observations (or groups of observations). The examples also show the important role of subject matter in guiding diagnostics, since judgement has to be made on how to deal with outliers or curvature. Diagnostics alone cannot decide the optimal model. Conversely, the diagnostics are quite helpful in focusing the experimenter’s attention to possible curvature, or to the critical few observations that need closer study, thus enabling knowledgeable decisions on the appropriate analysis, and the further enrichment of subject matter knowledge.

BIBLIOGRAPHY

Figure 5.7. For the Response Type Data, clockwise from the upper left corner: the Wilcoxon residual plot for the first-order model; the Wilcoxon residual plot for the second-order model; the GR residual plot for the first-order model; and the GR residual plot for the second-order model.


20


