INFLUENCE MEASURES FOR ROBUST REGRESSION

by

R. Dennis Cook and Sanford Weisberg

University of Minnesota

Technical Report No. 384

Department of Applied Statistics
School of Statistics
University of Minnesota
St. Paul, Minnesota 55108

15 December 1980

ABSTRACT

Methods for the detection of influential or important cases have been used and studied in several settings including linear least squares regression, logistic regression, and discriminant analysis. In this paper, analogous methods for assessing influence of individual cases in robust regression are proposed. Useful one-step (non-iterative) approximations are presented, and the limitations of these approximations are studied. This leads to definition of a second order diagnostic that has a large value when the one-step approximation is inadequate. Results are illustrated by application to three examples from the literature.

Keywords: Linear model, Distance measures, Leverage

1. Introduction

Influence, as measured by the theoretical influence curve, plays an important role in the study of various properties of estimators. The influence curve is used to measure the sensitivity of estimators to perturbations in the probability mechanism underlying observed data. The exposition of this idea by Hampel (1974) and its use in the Princeton study (Andrews et al., 1972) and in the subsequent literature had led to the study of classes of robust estimators; that is, estimators that are relatively insensitive to small perturbations in the underlying probability mechanism.

Recently, the idea of influence has been put to a related, but different use. Rather than study the effect of perturbing the underlying distribution, the observed data are perturbed, perhaps by deleting units or cases from the data. In this application, an empirical version of the influence curve is used to study the robustness of an estimation procedure on a particular data set. Thus, the notion of robustness is applied to a data set, rather than to an estimation method. Most of the development in this area has been for linear least squares regression (Andrews and Pregibon, 1978; Belsley, Kuh and Welsch, 1980; Cook, 1977, 1979; Cook and Weisberg, 1980; Hoaglin and Welsch, 1978; Johnson and Geisser, 1979), and for logistic regression (Pregibon, 1980), where one is led to consider diagnostic statistics that reflect the role or influence of cases or groups of cases on an analysis. The inclusion of some of these methods into widely distributed computer programs such as BMDP and MINITAB reflects their wide applicability and usefulness (e.g., BMDP9R, Dixon and Brown, 1979) and has greatly facilitated their use.

Generally, the diagnostics derived from the influence curve aid the analyst in detecting cases that seriously influence various aspects of an analysis, and in understanding and studying the structure of the data. For instance, influential cases tend to be associated with outlying responses or high leverage points (e.g., extreme points in the factor or design space).

Robust estimators are designed to reduce or bound the influence of outlying responses. For example, when robust methods are viewed as iteratively reweighted least squares, the weights resulting from a robust fit are useful for detecting outlying responses (Hogg, 1979); however, the weights will not in general be effective diagnostics for detecting influential cases arising because of high leverage. Robust estimates may be influenced more by leverage points than are least squares estimates. Hill (1977), Welsch (1977) and Krasker and Welsch (1979) have investigated various ways of modifying the usual robust estimators so that the influence of both outlying responses and high leverage points is bounded.

Using the sample influence curve (SIC), we extend the diagnostics developed for identifying influential cases in linear and logistic regression to robust regression. In Section 2, we give the general formulation and discuss the rationale behind the use of the SIC. A one-step approximation to the SIC is presented in Section 3 and is illustrated using the Huber ψ-function. In Section 4, the accuracy of the one-step approximation is illustrated using several data sets.

2. Robust Regression

Consider the usual linear regression model

$$Y = X\beta + e \tag{2.1}$$

where $Y = (y_i)$ is an n-vector of responses, $X = (x_{ij})$ is an $n \times p$ full rank matrix of known constants, β is a $p \times 1$ vector of unknown parameters, and $e = (e_i)$ is an n-vector of errors that are independent and identically distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. The vector (y_i, x_i^T) is called the ith case.

Under this formulation, a robust estimate $\tilde{\beta}$ of β is chosen to minimize

$$\sum_{i=1}^{n} \rho[(y_i - x_i^T\beta)/\tilde{\sigma}] \tag{2.2}$$

where ρ is a suitably selected loss function and $\tilde{\sigma}$ is a robust scale estimate that may be determined previously or simultaneously to achieve scale invariance. For a detailed discussion of robust regression and the associated computational methods see, for example, Huber (1977, p. 36) and Hogg (1979).

Let $\psi = \rho'$ and let $\Psi(\beta)$ denote an n-vector with elements $\psi_i(\beta) = \psi[(y_i - x_i^T\beta)/\tilde{\sigma}]$. For notational convenience, we set $\tilde{\Psi} = \Psi(\tilde{\beta})$ and $\tilde{\psi}_i = \psi_i(\tilde{\beta})$. A necessary condition for the minimization of (2.2) is that $\tilde{\beta}$ satisfy

$$X^T\tilde{\Psi} = 0. \tag{2.3}$$

If ρ is convex, this condition is also sufficient. Generally, (2.3) guarantees only that a local minimum has been achieved.

For illustration, the loss function proposed by Huber (1964, 1973) is given by

$$\rho(z) = \begin{cases} \frac{z^2}{2}, & |z| \leq c \\ c|z| - \frac{c^2}{2}, & |z| > c \end{cases} \tag{2.4}$$

where c is a selected positive constant. The corresponding Huber ψ-function is

$$\psi(z) = \begin{cases} -c, & z < -c \\ z, & -c \leq z \leq c \\ c, & z > c \end{cases} \tag{2.5}$$

The loss function corresponding to least squares is obtained from (2.4) by letting $c \to \infty$.

The influence of the ith case on a robust estimate $\tilde{\beta}$ can be determined by using the infinitesimal perturbation approach suggested by Belsley, Kuh and Welsch (1980) and Pregibon (1980). In this approach, the model (2.1) is modified by the specification $\mathrm{Var}(e_j) = \sigma^2$ for all $j \neq i$ and $\mathrm{Var}(e_i) = \sigma^2/w_i$, $0 \leq w_i \leq 1$. Robust estimation may then be applied to the transformed model,

$$W^{\frac{1}{2}}Y = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}e \tag{2.6}$$

where $W = \mathrm{diag}(w_j)$ and $w_j = 1$ for all $j \neq i$. Let $\tilde{\beta}(w_i)$ denote the robust estimator based on (2.6). The effects of infinitesimal perturbations in w_i can be seen by considering the rate of change in $\tilde{\beta}(w_i)$ with respect to w_i, with large values indicating that the ith case has a high influence on $\tilde{\beta}$. Generally,

$$\Delta\tilde{\beta}(w_i) = \frac{\partial}{\partial w_i}\,\tilde{\beta}(w_i) \tag{2.7}$$

will depend on w_i so that a complete understanding of the effects of perturbing w_i requires knowledge of the surface described by $\Delta\tilde{\beta}(w_i)$. Special cases are worthy of attention, however. First, evaluation of (2.7) at $w_i = 1$ describes the effects of small changes at $\tilde{\beta}$. The function $\Delta\tilde{\beta}(1)$ is essentially the empirical influence curve for $\tilde{\beta}$ (cf. Mallows, 1975; Pregibon, 1980). For diagnostic purposes, $\Delta\tilde{\beta}(1)$ is conservative and tends to ignore high leverage points, as is illustrated from the following least squares calculations. Let

$$V = (v_{ij}) = X(X^TX)^{-1}X^T \tag{2.8}$$

and

$$r = (r_i) = (I - V)Y. \tag{2.9}$$

Then for the least squares estimator $\hat{\beta}$ of β

$$\Delta\hat{\beta}(1) = (X^TX)^{-1}x_i r_i \tag{2.10}$$

(Belsley, Kuh and Welsch, 1980). Form (2.10) does not depend on v_{ii}, the usual measure of leverage (Cook and Weisberg, 1980).

Second, evaluation of (2.7) at $w_i = 0$ describes the change in $\tilde{\beta}$ as the ith case is deleted from the data. For diagnostic purposes, this tends to emphasize high leverage cases. For least squares,

$$\Delta\hat{\beta}(0) = (X^TX)^{-1}x_i r_i/(1 - v_{ii})^2 \tag{2.11}$$

Since v_{ii} is near 1 for high leverage cases, $\Delta\hat{\beta}(0)$ will tend to be large.

Figures 1 and 2, which derive from the cloud seeding data described by Cook and Weisberg (1980), serve to illustrate these remarks. Figure 1 is a graph of the β_{14} component (see equation (7.1) in Cook and Weisberg, 1980) of $\tilde{\beta}(w_2)$ against w_2. $\tilde{\beta}(w_2)$ was obtained by applying (2.4) with c = 1 to (2.6). First setting $w_2 = 1$, $\tilde{\beta}_{14}$ was computed via an iterative algorithm based essentially on Newton's method as described in Huber (1977, p. 38) and Holland and Welsch (1977). The value of w_2 was then decreased in steps to 0; at each step the last value of $(\tilde{\beta},\tilde{\sigma})$ was used as a starting value.

Figure 2 is the analogous plot for the weighted least squares estimate $\hat{\beta}(w_2)$ (Belsley, Kuh and Welsch, 1980; Pregibon, 1980). In both figures, the diagonal line connecting the estimates when $w_2 = 0$ and 1 is provided for reference. Clearly, both estimates of β_{14} are insensitive to perturbations near $w_2 = 1$, but are highly sensitive to perturbations when w_2 is small. For small values of w_2, the Huber estimate appears to be slightly more sensitive to perturbations than the least squares estimate. In either, however, evaluation at $w_2 = 0$ or 1 provides a misleading view of the influence of the second case on the estimate of β_{14}.

Following Cook and Wesiberg (1980) and Pregibon (1980), a useful compromise between these approaches is to use the average rate of change as a measure of influence. Since

$$\int_0^1 \Delta\tilde{\beta}(w_i)dw_i = \tilde{\beta}(w_i = 1) - \tilde{\beta}(w_i = 0) \qquad (2.12)$$

the average rate of change is simply the difference in the estimates computed with and without the case in question. Functions of the form given in (2.12) are called sample influence curves (SIC) by Devlin, Gnanadesikan and Kettenring (1975) and we adopt their terminology. For the

Figure 1. $\tilde{\beta}_{14}(w_2)$, Huber estimator

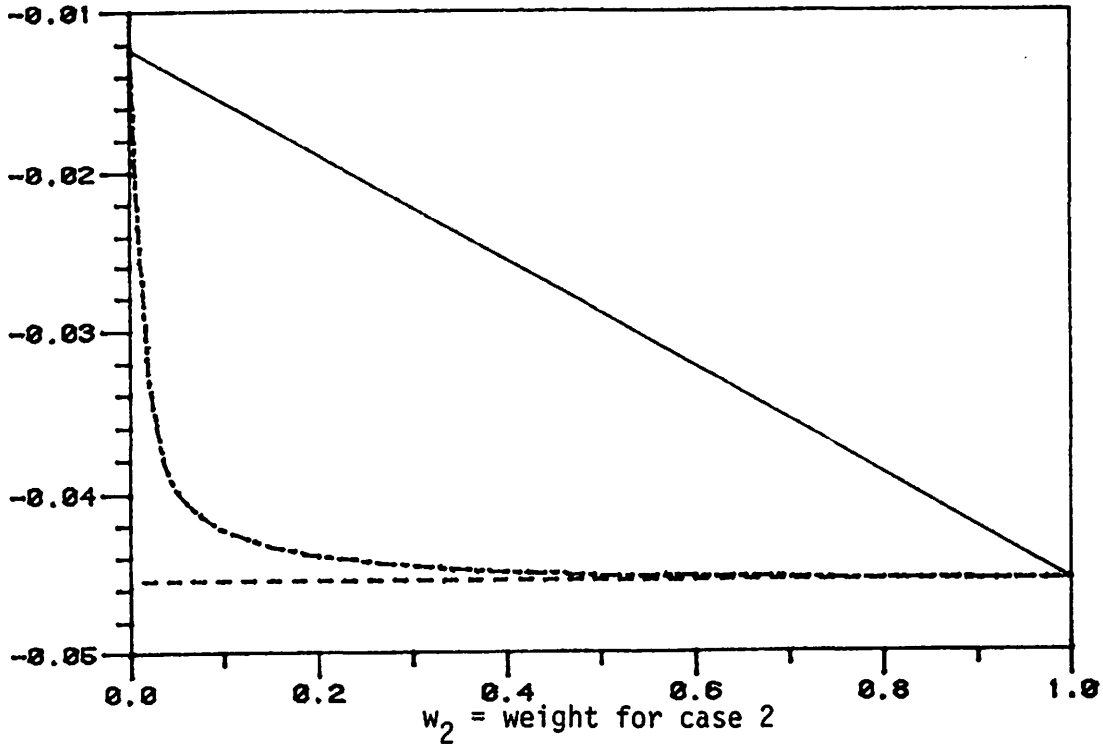
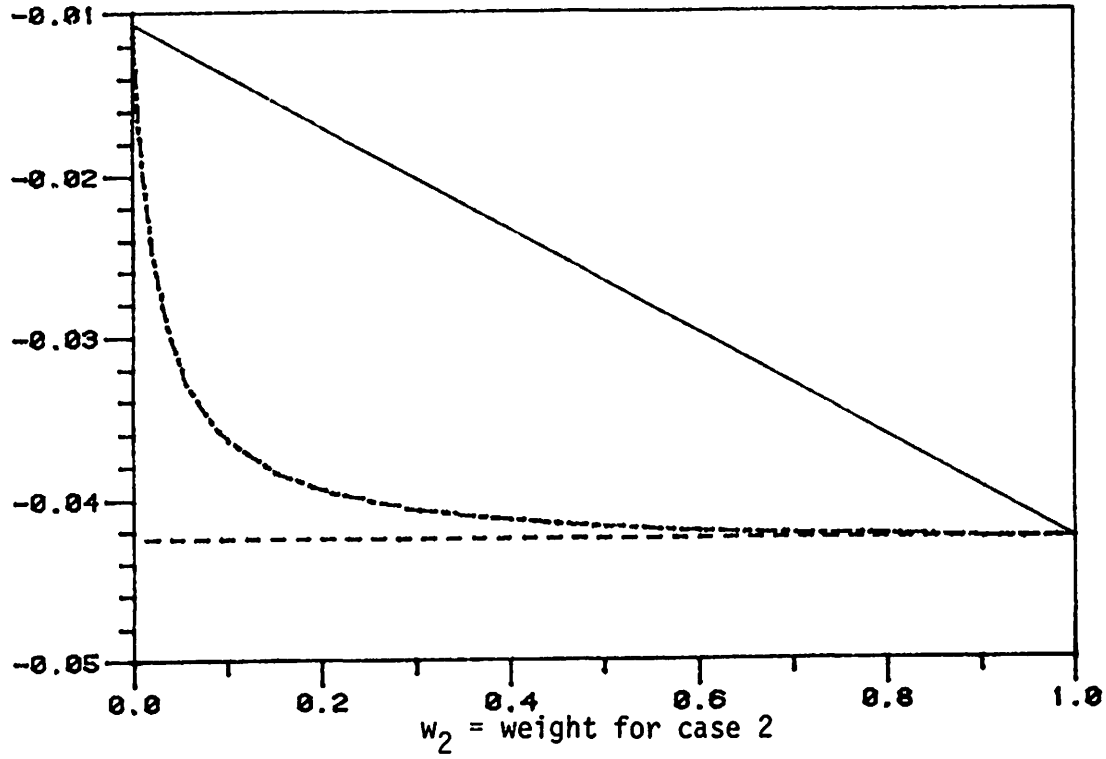


Figure 2. $\hat{\beta}_{14}(w_2)$, least squares estimator



β_{14} coefficient and $w_i = w_2$, $\tilde{\beta}_{14}(1) - \tilde{\beta}_{14}(0) = -.033$ while $\hat{\beta}_{14}(1) - \hat{\beta}_{14}(0) = -.032$. Thus, Huber's robust estimate of β_{14} is slightly more sensitive to case 2 than is the least squares estimate. However, case 2 does appear influential for both $\hat{\beta}_{14}$ and $\tilde{\beta}_{14}$.

Generally, we define the sample influence curve as

$$\delta_i(\tilde{\beta}) = \tilde{\beta} - \tilde{\beta}_{(i)} \tag{2.13}$$

where $\tilde{\beta}_{(i)}$ denotes the (robust) estimator of β computed without the ith case. Of course, determination of $\tilde{\beta}$ and $\tilde{\beta}_{(i)}$ requires an iterative scheme and thus the calculation of $\delta_i(\tilde{\beta})$, $i = 1,2,\ldots,n$, would require n+1 applications of the scheme. These calculations will be prohibitively expensive in general. In the next section, we suggest a one-step approximation to the sample influence curve. This approximation serves to reduce the amount of required calculation and to provide insight about characteristics of influential cases in robust regression. Also, since the SIC is intended as a diagnostic, there is no real need to find δ_i exactly as long as the diagnostic information based on the approximation is not seriously misleading. In addition, the one-step diagnostics are appropriate for the class of one-step robust estimators proposed by Bickel (1975), who provides asymptotic justification for their use.

3. A One-Step Sample Influence Curve

A one-step approximation to the SIC can be obtained by applying Newton's method (cf. Kennedy and Gentle, 1980, p. 442) to (2.3). Let $\tilde{\beta}$ and $\tilde{\sigma}$ denote fully iterated, robust estimates of location and scale, respectively, based on the full data. Further, let $\tilde{r}_i = (y_i - x_i^T\tilde{\beta})/\tilde{\sigma}$ and let $\tilde{P}$ denote an $n \times n$ diagonal matrix with ith diagonal element $\tilde{p}_i = \psi'(\tilde{r}_i)$.

Using $\tilde{\beta}$ and $\tilde{\sigma}$ as starting values, a single step using Newton's method (cf. Huber, 1977, p. 38; Holland and Welsch, 1977) applied to the data with the ith case removed gives $\tilde{\beta}_{(i)} \approx \tilde{\beta}^1_{(i)}$, where

$$\tilde{\beta}^1_{(i)} = \tilde{\beta} + \tilde{\sigma}(X^T_{(i)}\tilde{P}_{(i)}X_{(i)})^{-1}X^T_{(i)}\tilde{\Psi}_{(i)} \tag{3.1}$$

Here, $X_{(i)}$ and $\tilde{\Psi}_{(i)}$ are obtained by deleting the ith row from X and $\tilde{\Psi}$ respectively, and $\tilde{P}_{(i)}$ is obtained by deleting the ith row and column from $\tilde{P}$.

At least two other algorithms are available for determining robust estimates, as outlined by Huber (1977) and Holland and Welsch (1977). It is generally recognized, however, that Newton's method will yield the most accurate results for a fixed number of iterations. One disadvantage of Newton's method is that it requires ψ', although for the application here this does not seem to be a serious difficulty. A second and perhaps more important disadvantage of Newton's method is that the one-step approximation (3.1) cannot be guaranteed to decrease the objective function (2.2) if the Hessian matrix $(X^T_{(i)}\tilde{P}_{(i)}X_{(i)})$ is not positive definite. An indefinite Hessian matrix is possible when the ψ-function is redescending, as with Andrews' (1974) sine estimator. We shall assume that the Hessian matrix is positive definite; Kennedy and Gentle (1980, p. 443) discuss modifications of Newton's method for situations in which it is indefinite.

Using the usual updating expressions (cf. Cook and Weisberg, 1980), the one-step approximation of the SIC provided by (3.1) can be expressed more informatively in terms of the full data. Let

$$\tilde{v}_i = \tilde{p}_i x_i^T (X^T \tilde{P} X)^{-1} x_i. \tag{3.2}$$

Then the one-step approximation $d_i(\tilde{\beta})$ to $\delta_i(\tilde{\beta})$ can be expressed as

$$d_i(\tilde{\beta}) = \tilde{\beta} - \tilde{\beta}_{(i)}^1 = \frac{\tilde{\sigma}(X^T\tilde{P}X)^{-1}x_i\tilde{\psi}_i}{1 - \tilde{v}_i} \tag{3.3}$$

For least squares, $\psi(z) = z$, (3.3) is exact and reduces to

$$\delta_i(\hat{\beta}) = d_i(\hat{\beta}) = \frac{(X^TX)^{-1}x_i r_i}{1 - v_{ii}} \tag{3.4}$$

The correspondence between least squares and robust estimators should be clear from a comparison of (3.4) and (3.3). In particular, the residuals r_i in (3.4) have been replaced by the Winsorized residuals $\tilde{\sigma}\tilde{\psi}_i$ and the leverage values v_{ii} have been replaced by the analogous $\tilde{v}_i$.

3.1 Accuracy of the One-Step Approximation

The one-step approximation $d_i(\tilde{\beta})$ will be exact ($d_i(\tilde{\beta}) = \delta_i(\tilde{\beta})$) only if

$$X_{(i)}^T \Psi_{(i)}(\tilde{\beta}_{(i)}^1) = 0. \tag{3.5}$$

In general, and assuming that Newton's method would converge eventually to $\tilde{\beta}_{(i)}$ with $(\tilde{\beta},\tilde{\sigma})$ as starting values, the one step approximation $d_i(\tilde{\beta})$ will be close to $\delta_i(\tilde{\beta})$ if the left hand side of (3.5) is close to zero. For convenience, define $M = X^T\tilde{P}X$ and $\tilde{w}_{ij} = x_i^T M^{-1} x_j$, so that $\tilde{v}_{ii} = p_i\tilde{w}_{ii}$. Also, define

$$\tilde{\varepsilon}_{ij} = \frac{\tilde{w}_{ij}}{1 - \tilde{v}_{ii}} \tilde{\psi}_i \tag{3.6}$$

with $\tilde{\psi}_i$ the ith element of $\tilde{\Psi}$. Using (3.3), the jth element of $\Psi_{(i)}(\tilde{\beta}^1_{(i)})$ can be written as

$$\psi[(y_j - x_j^T \tilde{\beta}^1_{(i)})/\tilde{\sigma}] = \psi[\tilde{r}_j + \tilde{\varepsilon}_{ij}] \tag{3.7}$$

Since $X^T\tilde{\Psi} = 0$, a sufficient condition for (3.5) to hold is $\tilde{\varepsilon}_{ij} = 0$, $j = 1,2,\ldots,n$, which will happen, for example, if $\tilde{\psi}_i = 0$. This will occur with some frequency with redescending estimators, such as Andrews' sine. Also, for the redescenders, $\tilde{\beta}_{(i)} = \tilde{\beta}$, and $\delta_i(\tilde{\beta}) = 0$, if $\tilde{\psi}_i = 0$.

If $|\tilde{\varepsilon}_{ij}|$ is sufficiently small, (3.7) can be expanded in Taylor series about 0 to give

$$\psi(\tilde{r}_j + \tilde{\varepsilon}_{ij}) = \psi(r_j) + \tilde{\varepsilon}_{ij}\psi'(\tilde{r}_j) + R_{ij} \tag{3.8}$$

where $R_{ij} = O(\tilde{\varepsilon}^2_{ij})$. In matrix notation, (3.8) may be rewritten as

$$\Psi_{(i)}(\tilde{\beta}^1_{(i)}) = \tilde{\Psi}_{(i)} + \frac{\tilde{P}_{(i)}X_{(i)}\tilde{M}^{-1}x_i\tilde{\psi}_i}{1 - \tilde{v}_i} + R_i \tag{3.9}$$

where R_i is an n-vector where the jth coordinate is R_{ij}. Multiplying (3.9) on the left by $X^T_{(i)}$ and simplifying leads to

$$\begin{aligned} X^T_{(i)}\Psi_{(i)}(\tilde{\beta}^1_{(i)}) &= X^T\tilde{\Psi} + X^T_{(i)}R_i \\ &= X^T_{(i)}R_i. \end{aligned} \tag{3.10}$$

Thus, the left side of (3.5) is $O(\hat{\varepsilon}^2_{ij})$ when $|\hat{\varepsilon}_{ij}|$ is sufficiently small for the Taylor series to be valid.

When ψ is piecewise linear (e.g., Huber's estimator), all derivatives of ψ beyond the first are 0, and hence $R_i = 0$. Thus, when-

ever the $|\tilde{\varepsilon}_{ij}|$ are sufficiently small, Newton's method will converge in one step. In particular, if the classification of $\tilde{r}_j$ according to the pieces of ψ is the same as the classification of the one step residuals $\tilde{r}_j + \tilde{\varepsilon}_{ij}$ then Newton's method will converge in one step.

Since the $|\tilde{\varepsilon}_{ij}|$ reflect the probable accuracy of the one-step approximation, they may be used to develop second order diagnostics that indicate when further iteration may be necessary. The second order diagnostic $\max_{j\neq i} |\tilde{\varepsilon}_{ij}|$ is an attractive candidate. Small values of this statistic indicate that the one-step approximation is probably sufficiently accurate for case i, while "large" values indicate that further iteration may be necessary. Unfortunately, the cost of computing this diagnostic will be roughly the same as that of a second iteration. When $\tilde{M}$ is positive definite, $\max_{j\neq i} |\tilde{\varepsilon}_{ij}|$ is bounded above for i = 1,2,...,n, by

$$b_i = \frac{|\tilde{\psi}_i|}{1 - \tilde{v}_i} (\tilde{w}_{ii} \max_{j\neq i} \tilde{w}_{jj})^{\frac{1}{2}} \qquad (3.11)$$

These bounds provide less costly alternative second order diagnostics. Comments on the use and effectiveness of these bounds are provided in Section 4.

3.2 Interpretation of the One-Step Approximation

The standard robust estimators reduce or bound the influence of cases corresponding to large residuals, but may be insensitive to high leverage cases. Huber (1977) comments that robust regression may not work well when highly influential cases are present.

Assuming a sufficiently accurate one-step approximation, the effects of leverage on robust estimators can be illustrated by using (3.3) in combination with Huber's ψ-function. In this case, the diagonal elements of P are

$$\tilde{p}_i = \begin{cases} 1, & |\tilde{r}_i| \leq c \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

and thus,

$$\tilde{v}_i = \begin{cases} x_i^T \tilde{M}^{-1} x_i = \tilde{w}_{ii}, & |\tilde{r}_i| \leq c \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

The one-step SIC, $d_i(\tilde{\beta})$, can now be expressed as

$$d_i(\tilde{\beta}) = \begin{cases} -\tilde{\sigma}\tilde{M}^{-1} x_i c, & \tilde{r}_i < -c \\ \dfrac{\tilde{\sigma}\tilde{M}^{-1} x_i \tilde{r}_i}{1 - \tilde{w}_{ii}}, & -c \leq \tilde{r}_i \leq c \\ \tilde{\sigma}\tilde{M}^{-1} x_i c, & \tilde{r}_i > c \end{cases} \tag{3.14}$$

If $-c < \tilde{r}_i < c$, the influence due to leverage can be expected to be greater than for least squares. This follows for Huber's ψ-function since $v_{ii} \leq \tilde{w}_{ii}$, $i = 1,2,\ldots,n$. Similarly, when $|\tilde{r}_i| > c$ the influence is less than that for least squares. Consider, for example, the situation in which $\tilde{v}_i = 0$ but, for all $k \neq i$, $\tilde{v}_k = 1$. If $\tilde{r}_i$ is positive,

$$d_i(\tilde{\beta}) = \frac{(X^TX)^{-1}x_i}{1 - v_{ii}} \quad (\tilde{\sigma c})$$

$$= \delta_i(\hat{\beta}) \frac{\tilde{\sigma c}}{r_i}$$

which is the SIC for least squares (3.4) reduced by the factor $\tilde{\sigma c}/r_i$.

4. Illustrations

Judgments about relative influence require a comparison of the p-dimensional vectors $d_i(\tilde{\beta})$, $i = 1,2,\ldots,n$ (or, if available, $\delta_i(\tilde{\beta})$). Any of the methods for comparing multidimensional vectors, such as Andrews' (1972) plots or Wilks' (1963) outlier detection criterion, can be used. Gnanadesikan (1977) discusses these and other appropriate techniques.

For linear least squares, past investigations (Belsley, Kuh and Welsch, 1980; Cook, 1977, 1979; Cook and Weisberg, 1980) indicate that the length of the SIC relative to a selected metric will give the essential diagnostic information. For robust regression, we consider two distance measures. The first $D(\delta_i(\tilde{\beta}))$ is based on the asymptotic covariance matrix of $\tilde{\beta}$,

$$D(\delta_i(\tilde{\beta})) = K \frac{\delta_i^T(\tilde{\beta})X^TX\delta_i(\tilde{\beta})}{p\tilde{\sigma}^2} \tag{4.1}$$

where K is a correction factor that depends on the ψ-function (see, for example, Hogg, 1979; Huber, 1977). For diagnostic purposes, K may be unimportant.

The second distance measure $D_W(\delta_i(\tilde{\beta}))$ is given by

$$D_W(\delta_i(\tilde{\beta})) = \frac{\delta_i^T(\tilde{\beta})X^TWX\delta_i(\tilde{\beta})}{p\tilde{\sigma}^2} \tag{4.2}$$

where $W = \text{diag}\{\psi(\tilde{r}_i)/\tilde{r}_i\}$ (W is not the same as in Section 2). This measure is based on using weighted least squares to obtain an approximation of the covariance matrix for $\tilde{\beta}$. Useful alternative interpretations of (4.1) and (4.2) can be obtained by reexpressing them in terms of the vectors of fitted values, $X\tilde{\beta}$ and $X\tilde{\beta}_{(i)}$. Norms other than (4.1) and (4.2) can be constructed to reflect more specific concerns (Cook and Weisberg, 1980).

A case will be called influential if (4.1) or (4.2) is large. In least squares regression, judgment of size of the normed measure is available by analogy to confidence intervals. For robust methods, where the metrics refer to approximate confidence contours, interpretation of size of the measure is more difficult. Generally, however, we will consider the values of the normed measure relative to the other values in a particular problem. A rough cut-off of 1.0 for either norm may be recommended, although further work on this issue remains to be done.

At the ith case, the one-step approximation $d_i(\tilde{\beta}) = d_i$ will be judged to provide the same diagnostic information as $\delta_i(\tilde{\beta}) = \delta_i$ whenever $D(d_i) \approx D(\delta_i)$ and/or $D_W(d_i) \approx D_W(\delta_i)$. This approach ignores the angle between δ_i and d_i, but knowledge about this angle seems inessential since routine diagnostic information must typically be based on a norm. Of course, once an influential case is found using $D(d_i)$ or $D_W(d_i)$, additional iterations can be applied for a more accurate determination of δ_i.

Some care must be exercised when using D or D_W if X^TX or X^TWX have large eigenvalues. For example, if X^TX has a single large eigenvalue and δ_i lies in the direction of the corresponding eigenvector, then $D(d_i)$ and $D(\delta_i)$ can be very different even if d_i and δ_i are "close". The implicit distance measure or norm in Newton's method is based on the

inner-product matrix $X_{(i)}^T \tilde{P}_{(i)} X_{(i)}$ which is reflected more closely by X^TWX than by X^TX. For this reason we expect that $|D_W(\delta_i) - D_W(d_i)|$ will generally be smaller than $|D(\delta_i) - D(d_i)|$. In the examples to be considered here, results using (4.1) are qualitatively similar to the results for (4.2) except that the one-step approximations are on average somewhat worse. For brevity, we will discuss only (4.2) in what follows.

The accuracy of d_i was investigated by examining both the fully iterated influence values and the one-step approximation for a number of different data sets having different X-structures and for two estimators. The estimators used were Huber's (2.4) with tuning constant c = 1.345 and Andrews' sine estimator given by

$$\psi(z) = \begin{cases} \sin(z/k), & |z| \leq k\pi \\ 0, & |z| > k\pi \end{cases} \qquad (4.3)$$

with tuning constant k = 1.5. In each case, $\tilde{\beta}_{(i)}$ was computed by 10 steps of an iterative procedure based on Bickel's (1975) proposal 2 with $(\tilde{\beta},\tilde{\sigma})$ as starting values. Thus, the computed values of δ_i may not correspond exactly to a stationary point of (3.5). However, the full data estimate $\tilde{\beta}$ was computed with enough iterations to reach a stationary point of (2.3) with at least 7 digits of accuracy in each coefficient.

4.1 Stack Loss Data

Daniel and Wood (1980, p. 61) give a set of data usually called the "stack loss" data, with n = 21, p = 3. We have fit the first order model as in Andrews (1974). Figure 3 gives a summary of the influence analysis using D_W defined by (4.2) for the sine estimator. This is a plot of b_i (see, (3.11))

Figure 3. Sine estimator, stack loss data

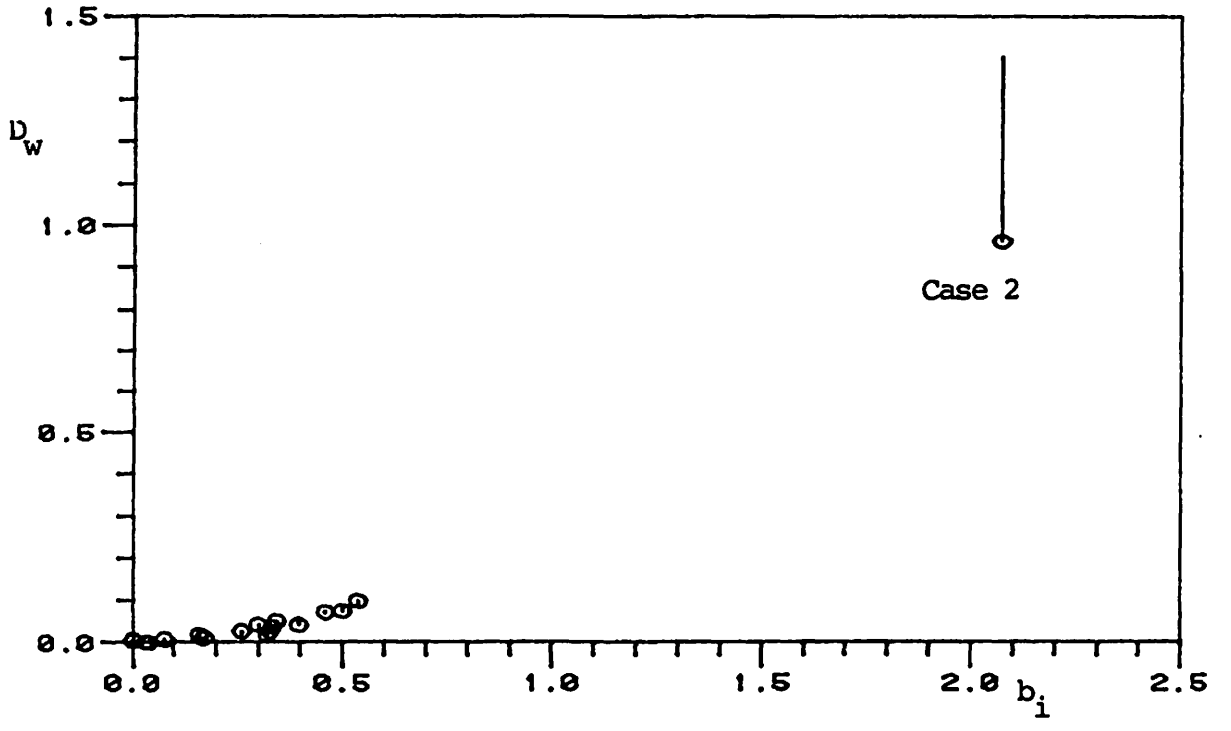
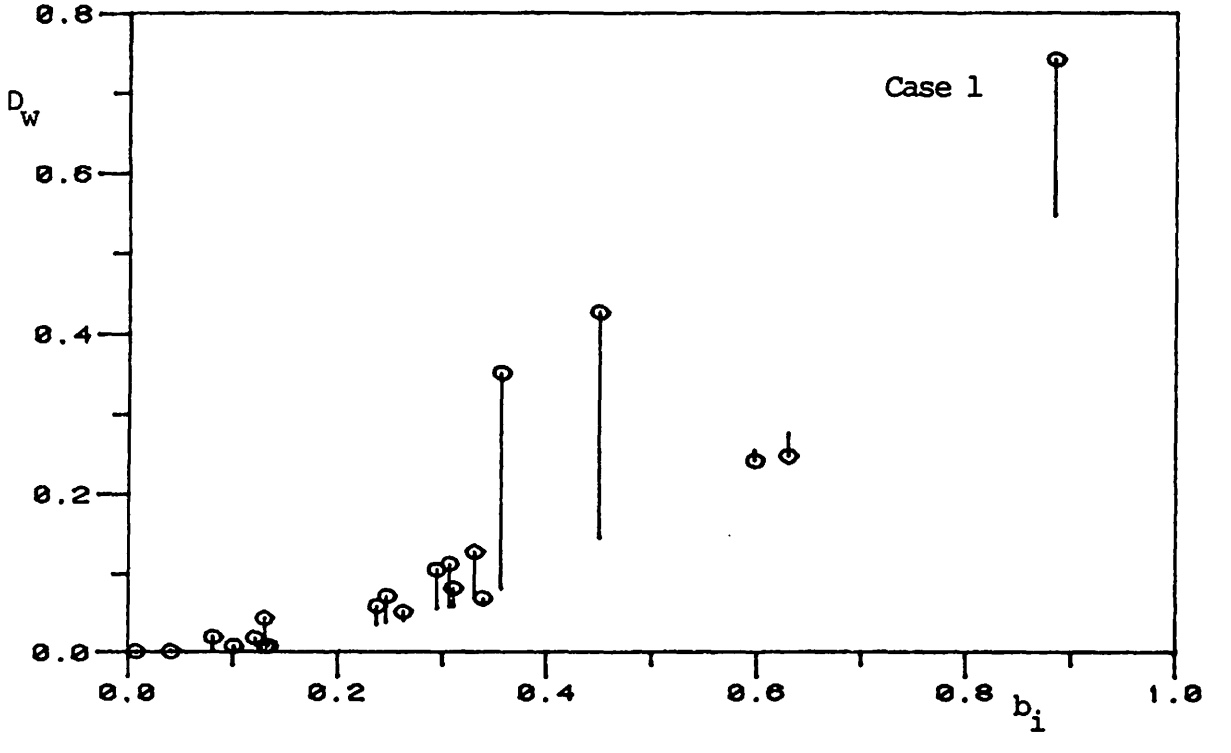


Figure 4. Huber estimator, stack loss data



on the x axis versus both $D_W(d_i)$ and $D_W(\delta_i)$. Values of $D_W(\delta_i)$ are indicated by small circles and the corresponding $D_W(d_i)$ is at the end of the line segment starting at the center of the circle. Long line segments are indicative large values of $|D_W(\delta_i) - D_W(d_i)|$. Line segments pointing downward correspond to cases with influence underestimated by the one-step approximation; while upward pointing cases correspond to overestimates. For the sine estimator in Figure 3, the one-step approximation is excellent: only one case, number 2, with $b_2 = 2.07$, has a one-step approximation that is much different from the fully iterated value. Since b_2 was relatively large, additional iteration for this case was indicated. The value of $D_W(\delta_2) = 0.96$ indicates that case 2 was in fact influential for estimating $\tilde{\beta}$.

Figure 4 gives the equivalent plot for the Huber estimator. In this graph, it is apparent that the magnitude of $|D_W(d_i) - D_W(\delta_i)|$ increases with b_i. However, since the maximum b_i is $b = 0.88$, no further iteration (or at most, further iteration for case 1) would be required. For this estimator, no cases are overly influential as the maximum $D_W(\delta_i) = D_W(\delta_1) = 0.74$. Also, by comparing Figures 3 and 4, we see that the most influential case for the sine estimator is uninfluential for the Huber estimator and <u>vice versa</u>. Indeed, for the sine estimator, $\tilde{\psi}_1 = 0$ and hence $D_W(\delta_1(\tilde{\beta})) = 0$. Thus a case that is influential for one estimator need not be influential for another.

4.2 Cloudseeding

Cook and Weisberg (1980) and Weisberg (1980) present a data set on cloudseeding with $n = 24$ and $p = 11$. For the sine estimator (Figure 5), the values of the second order diagnostic are very large: b_i exceeds 1 on 17 of 24 cases, and the one-step approximation cannot be expected to be

Figure 5. Sine estimator, cloudseeding data

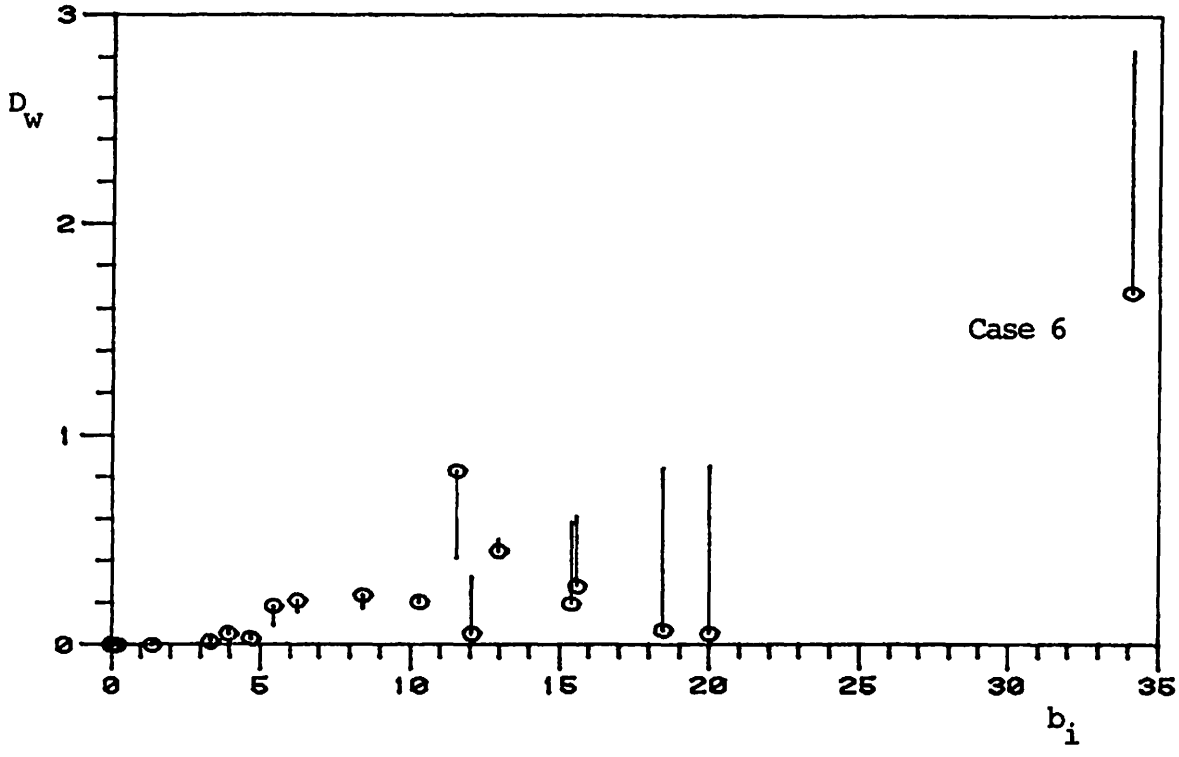
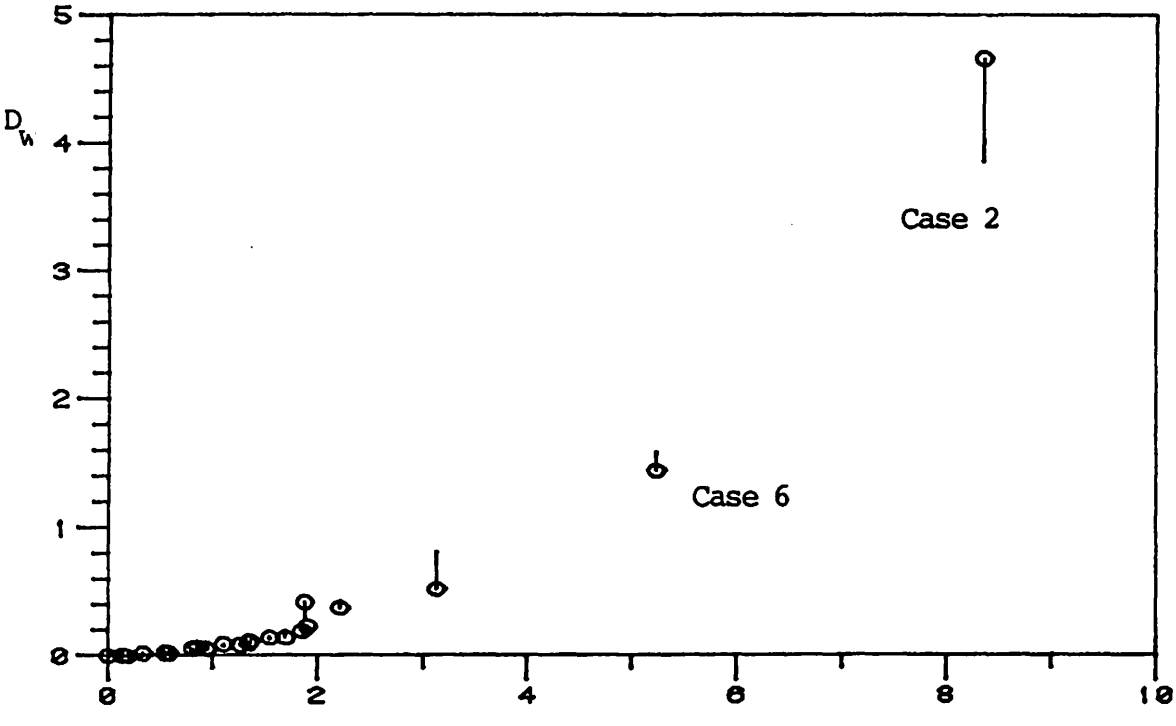


Figure 6. Huber estimator, cloudseeding data



an accurate measure of influence. As shown in Figure 5 for the sine estimator, both large overestimates and underestimates are apparent. For this data set and the sine estimator, further iteration would be required to assess influence. However, from the $D_w(\delta_i)$ values (the circles), we see that only one case, number 6, is clearly influential for these data.

For the Huber estimator, Figure 6, the results are similar. Here, 13 b_i exceed 1, although from the figure it is clear that the one-step estimators are adequate. Case 2 with $D_w(\delta_2) = 4.65$ would be judged to be very influential, as was found by Cook and Weisberg (1980) for the least squares estimator.

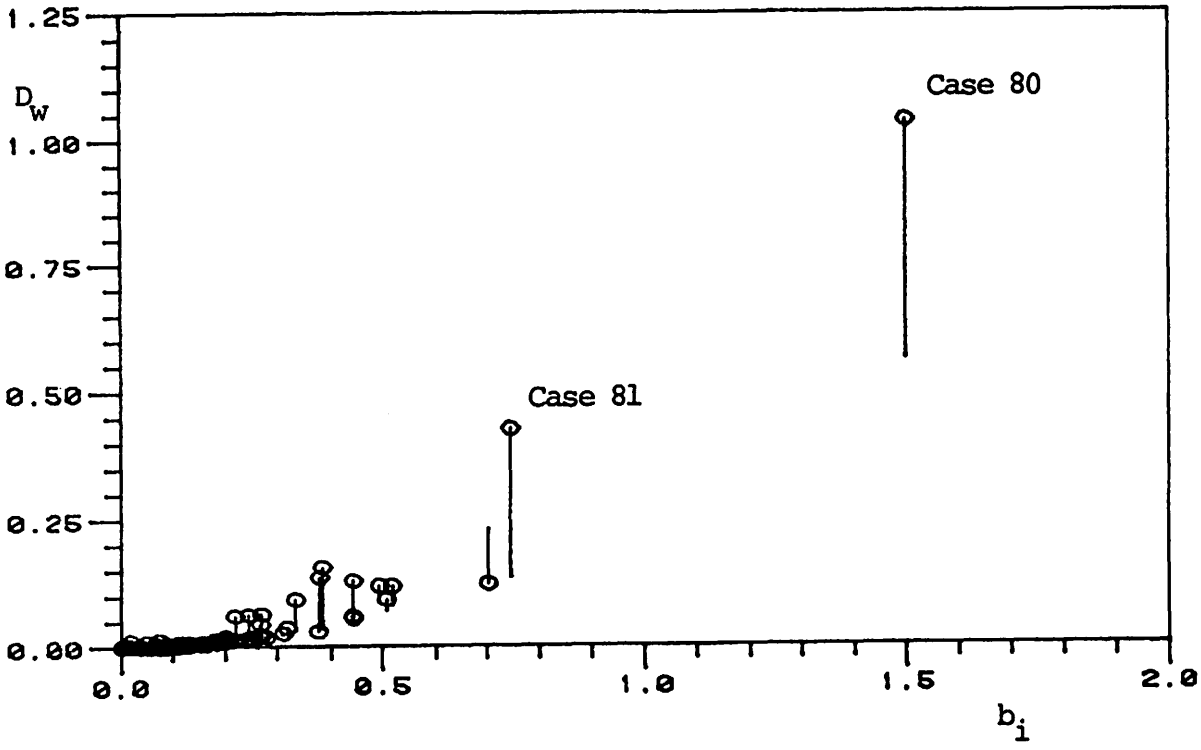
4.3 A Larger Data Set

Krasker and Welsch (1980) provide an economic data set with $n = 84$ and $p = 9$. For the Huber estimate, the influence is summarized by Figure 7. It is clear from this plot that further iteration is required only for one case, case 80, with $b_{80} = 1.45$, since all the other b_i are smaller than 0.75. Full iteration for case 80 gives $D(\delta_{80}) = 1.04$ indicating that this case is in fact influential. In most data sets with sample sizes as large or larger than that in this example, the number of large b_i can be expected to be relatively small.

As suggested by these examples, we expect the one-step approximation to be sufficiently accurate whenever b_i is less than about 1/2 of the tuning constant. Full iteration is a wise precaution whenever b_i is larger than the tuning constant. In intermediate situations, one or two additional steps may be sufficient.

As mentioned in Section 3, two other algorithms are available for

Figure 7. Huber estimator, economic data



determining robust estimates. In these examples, the one-step approximations based on the alternative algorithms discussed by Huber (1977) proved to be considerably less accurate than the one-step approximation used here.

5. Acknowledgements

The computations reported in this paper were carried out using a program for robust regression written by Christopher Bingham, with additional code for the influence calculations added by Norton Holshuh. This work was supported in part by a grant from the National Institute of General Medical Sciences.

References

Andrews, D.F. (1972), "Plots of high-dimensional data," Biometrics 28, 125-136.

Andrews, D.F. (1974), "A robust method for multiple linear regression," Technometrics 16, 523-531.

Andrews, D.F., P. Bickel, F. Hampel, P. Huber, W. Rogers and J. Wm. Tukey (1972), Robust Estimates of Location, Princeton, N.J.: Princeton University Press.

Andrews, D.F. and D. Pregibon (1978), "Finding outliers that matter," Journal of the Royal Statistical Society, Series B, 40, 85-93.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), Regression Diagnostics, New York: Wiley.

Bickel, P.J. (1975), "One-step Huber estimates in the linear model," Journal of the American Statistical Association, 70, 428-434.

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

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

Cook, R.D. and S. Weisberg (1980), "Characterizations of an empirical influence function for detecting influential curves in regression," Technometrics, in press.

Daniel, C. and F. Wood (1980), Fitting Equations to Data, New York: Wiley, Second Edition.

Devlin, S.J., R. Gnanadesikan and J.R. Kettenning (1975), "Robust estimates and outlier detection with correlation coefficients," Biometrika, 62, 531-546.

Dixon, W. and M. Brown (1979), BMDP-79, Los Angeles, University of California.

Gnanadesikan, R. (1977), Methods for Statistical Analysis of Multivariate Data, New York: Wiley.

Hampel, F. (1974), "The influence curve and its role in robust estimation," Journal of the American Statistical Association, 69, 383-393.

Hill, R.W. (1977), "Robust regression when there are outliers in the carriers," Ph.D. Dissertation, Harvard University.

Hoaglin, D.C. and R. Welsch (1978), "The hat matrix in regression and ANOVA," American Statistician, 32, 17-22.

Hogg, R.V. (1979), "Statistical robustness: One view of its use in applications today," American Statistician, 33, 108-115.

Holland, P. and R. Welsch (1977), "Robust regression using iteratively reweighted least squares," Communication in Statistics, A6, 813-828.

Huber, P. (1964), "Robust estimation of a location parameter," Annals of Mathematical Statistics, 35, 73-101.

Huber, P. (1973), "Robust regression: Asymptotics, conjectures, and Monte Carlo," American Statistician, 1, 799-821.

Huber, P. (1977), Robust Statistical Procedures, No. 27, Regional Conference Series in Applied Mathematics. Philadelphia: Society for International and Applied Mathematics.

Johnson, Wesley and S. Geisser (1979), "Assessing the predictive influence of observation," School of Statistics Technical Report No. 355, University of Minnesota.

Kennedy, W. and J. Gentle (1980), Statistical Computing, New York: Dekker.

Krasker, W.S. and R.E. Welsch (1979), "Efficient bounded-influence regression estimation using alternative definitions of sensitivity," Technical Report No. 3, Center for Computational Research in Economics and Management, Massachusetts Institute of Technology.

Mallows, C.L. (1975), "On some topics in robustness," unpublished, Bell Telephone Laboratories, Murray Hill, N.J.

Pregibon, D. (1980), "Logistic regression diagnostics," Annals of Statistics, in press.

Weisberg, S. (1980), Applied Linear Regression, New York: Wiley.

Wilks, S.S. (1963), "Multivariate statistical outliers," Sankhyā, 25, 407-426.

Welsch, R. (1977), "Regression sensitivity analysis and bounded-influence regression," forthcoming in Evaluation of Econometric Models, J. Kmenta and J. Ramsey, eds., New York: Academic Press.