ROBUST REGRESSION ESTIMATORS BASED ON DELETED RESIDUALS

Robert K. Henderson
Engineering Department
E.I. DuPont de Nemours and Co.
Wilmington, DE 19898

and

Richard F. Gunst
Department of Statistics
Southern Methodist University
Dallas, TX 75275

ABSTRACT

Robust regression estimators guard against distortion of parameter estimates because of outliers or certain violations of error assumptions. Robust M-estimators which are direct extensions of M-estimation for location models are not resistant to a small number of extreme data values in a regression analysis. In this paper a general form for a regression influence function is derived and related to other theoretical and sample-based regression influence functions which have been studied. Expressions for the influence function suggest alternative M-estimators based on deleted residuals rather than raw residuals. Several alternative estimators are proposed and compared on two data sets.

Key Words: Bounded-Leverage; Influence Functions; M-Estimation

1. INTRODUCTION

Robust regression estimators are intended to provide protection from a variety of model assumption violations and data anomalies, including heavy-tailed or contaminated error distributions and extreme or "high-leverage" predictor-variable values. A growing concern (and stimulus for much current research) is the realization that an estimator which is robust to violations of the usual error assumptions might not be resistant to changes in a small number of predictor-variable values and vice-versa. In particular, there is a need for procedures which will enable one to detect a small number of unusual or discrepant observations (y_i, x_i') and isolate the effects of these observations on an estimator. One important measure in such an assessment is the influence function. This paper investigates the role of regression influence functions and sample-based approximations to influence functions. As a result of this investigation several new robust regression estimators are suggested and their potential for effective use in the presence of unusual observations is studied.

Influence functions were introduced by Hampel (1968, 1971, 1974) as quantitative measures of the sensitivity of estimators to violations of model assumptions. They have proven especially useful for studying local robustness properties of estimators; e.g., for studying the effects of extreme data values on location estimators for univariate symmetric populations. Because of the sensitivity of least squares estimators to extreme observations, influence functions are potentially of great value in studying the effects of extreme observations on least squares and robust regression estimators.

Regression influence functions have been defined by Hinkley (1977), Krasker and Welsch (1982), Dorsett and Gunst (1982) and others. Hinkley used an examination of the influence function to suggest a weighted, jackknifed (least squares) estimator in which the difference between the least squares estimate and the ith weighted pseudo-value equals an estimated influence function of β at (y_i, x_i'). Krasker and Welsch derive a weighted least squares estimator which is asymptotically efficient among weighted least squares estimators which bound the gross-error sensitivity of the influence function and whose weights depend on y only as a function of $|y-x'\beta|$. Dorsett and Gunst use the influence function to suggest a weighting of predictor variable values to achieve bounded leverage, followed by M-estimation on the residuals to protect against extreme response values.

As one might expect, the influence functions derived by the above authors can be shown to be equivalent under the same model assumptions; however, there are many sample-based influence functions which have appeared in the literature. Mallows (1975), Belsley, Kuh and Welsch (1980) and Cook and Weisburg (1980, 1982) discuss many of the sample-based versions which have been proposed. Due to computational complexities which arise from simultaneously measuring the effect of several extreme observations, most of the sample-based influence functions which are readily computable only measure the effect of adding or deleting a single observation. A few proposals for treating several extreme observations appear in Andrews and Pregibon (1978), Dempster and Gasko-Green (1981) and Draper and John (1981). Because theoretical derivations and computational procedures for handling single extreme observations are so much better refined than for multiple observations, we will concentrate attention on robust estimation for this case; nevertheless, comments on generalizations of the following results will be made when appropriate.

The outline of this paper is as follows. Section 2 defines a general influence function for multiple linear regression models. Several influence functions which have been discussed in the literature are shown to be special cases of this general influence function. In Section 3 a few of the more popular sample-based influence functions are discussed. From the results of Sections 2 and 3, suggestions for alternative robust regression estimators emerge. These estimators are discussed and are used to analyze two data sets in Section 4. Concluding remarks are made in Section 5.

2. REGRESSION INFLUENCE FUNCTIONS

A regression functional $T_\beta(F,\psi)$ can be defined implicitly from the estimating equation

$$\int\cdots\int X'\Psi(Y-XT_\beta(F,\psi))dF_X(Y) = 0 \qquad (2.1)$$

where $\Psi(t)' = (\psi(t_1),\ldots,\psi(t_n))$ for some specified function $\psi(\cdot)$. The distribution function $F_X(Y)$ is understood to be the conditional distribution of Y given X and can be obtained from the joint distribution of (y,x') or from the usual regression model assumptions:

$$Y = X\beta + \varepsilon \quad , \tag{2.2}$$

where Y is an n-dimensional vector of response variables, X is an $(n \times m)$ full-column-rank matrix of nonstochastic predictor variables, β is an m-dimensional vector of unknown regression coefficients, and ε is an unobservable n-dimensional vector of error variables with $\varepsilon_i \sim NID(0,\sigma^2)$. If $\psi(t_i) = t_i$, corresponding to least squares estimation, and $dF_X(Y)$ is the assumed multivariate normal density function from model (2.2), then $T_\beta(F,\psi_{LS}) = \beta$.

The functional equation (2.1) allows a great deal of flexibility for examining different estimators (through the choice of $\psi(\cdot)$) and error distributions (through the choice of $F_X(\cdot)$) in the study of robust regression estimators. For example, an alternative to model (2.2) which is of special interest in the sequel assumes that n observations correspond to model (2.2) but an additional k observations are point masses at (Z,U). In this case,

$$T_\beta(F,\psi_{LS}) = (X_+'X_+)^{-1}(U'Z + X'X\beta) \tag{2.3}$$

where $X_+' = (X':U')$. Equation (2.3), which is readily reducible to a single point mass at (z,u'), shows that parametric values (hence, estimators) can be affected by both extreme response and extreme predictor variable values through the individual elements of Z and U. Although not studied in detail in this article, the conditional assumption on X can be relaxed and the dependence of the regression estimator on the distribution of x' can be studied by replacing

$dF_X(Y)$ with $dF(Y,X)$ in equation (2.1), similar to the approach of Krasker and Welsch (1982).

An influence function for the estimator $T_\beta(F,\psi)$ is defined as

$$T_\beta(F,G,\psi) = \lim_{\alpha\to 0^+}\{T_\beta(F^\alpha,\psi)-T_\beta(F,\psi)\}/\alpha \ , \qquad (2.4)$$

where F^α represents a contaminated distribution of the form

$$F^\alpha(Y) = (1-\alpha)F(Y) + \alpha G(Y) \qquad 0 \leq \alpha \leq 1. \qquad (2.5)$$

In the expression for the contaminated distribution (2.5), $F(Y) = F_X(Y)$ and $G(Y)$ is an alternative conditional distribution for the response variables. The study of outliers is facilitated by allowing this latter conditional distribution of Y to be based on an alternative, perhaps completely distinct, choice of X, say X_0. Thus $G(Y) = G_{X_0}(Y_0)$ is defined to be a distribution function corresponding to the model

$$Y_0 = X_0\beta + \varepsilon_0 \ , \qquad (2.6)$$

where Y_0 is an n_0-dimensional response vector and the distribution of some or all of the elements of ε_0 might differ from the assumed $NID(0,\sigma^2)$. Allowing $n_0 \neq n$ causes no theoretical difficulties in any of the subsequent results since equal sample sizes can always be obtained by augmenting the distribution of Y or Y_0 by degenarate distributions at $(y_i,x_i') = (0,0')$.

If the limit in equation (2.4) exists and is unique from positive and negative directions the influence function for $T_\beta(F,\psi)$ can be obtained by differentiation of

$$(1-\alpha)\int\cdots\int X'\Psi(Y-XT_\beta(F,\psi))dF_X(Y)$$
$$+ \alpha\int\cdots\int X_0'\Psi(Y_0-X_0T_\beta(F,\psi))dG_{X_0}(Y_0) = 0. \qquad (2.7)$$

The influence function so obtained is

$$I_\beta(F,G,\psi) = (X'\dot{D}_\psi X)^{-1}\int\cdots\int X_0'\Psi(Y_0 - X_0 T_\beta(F,\psi))\,dG_{X_0}(Y_0) \;, \qquad (2.8)$$

where

$$\dot{D}_\psi = \text{diag}\{\int\dot{\psi}(y_i - x_i' T_\beta(F,\psi))\,dF_{x_i}(y_i)\}$$

and $\dot{\psi}(t_i) = d\psi(t_i)/dt_i$.

2.1 Special Cases

Let $G_{X_0}(Y_0)$ denote a point-mass distribution at Y_0. Then

$$I_\beta(F,G,\psi) = (X'\dot{D}_\psi X)^{-1} X_0'\Psi(Y_0 - X_0 T_\beta(F,\psi)) \qquad (2.9)$$

is the influence function studied by Dorsett and Gunst (1982). By letting $X_0 = u'$, $Y_0 = z$, and appropriately choosing the function $\psi(\cdot)$, the bounded-influence estimator of Krasker and Welsch (1982) can be shown to have an influence function similar to (2.9). Least squares estimation corresponds to $\psi(t_i) = t_i$. The least squares influence function corresponding to a point-mass contaminant at Y_0 is

$$I_\beta(F,G,\psi_{LS}) = (X'X)^{-1} X_0'(Y_0 - X_0\beta) \;; \qquad (2.10)$$

further, if $Y_0 = z1$ and $X_0 = 1u'$ for some n-dimensional row vector u', where 1 is an n-dimensional vector of ones, then

$$I_\beta(F,G,\psi_{LS}) = n(X'X)^{-1} u(z - u'\beta) \;, \qquad (2.11)$$

which is the influence function used by Hinkley (1977). Note that if $E[\varepsilon_0] = 0$ in equation (2.6) then $I_\beta(F,G,\psi_{LS}) = 0$, regardless of whether $X = X_0$, since the integral in equation (2.8) is zero.

Consider now an approach suggested by some of the sample-based influence functions discussed in Mallows (1975), Belsley, Kuh and Welsch (1980), and Cook and Weisberg (1982). Let $X_0' = (X':U')$,

$Y_0' = (Y':Z')$, and suppose that the first n elements of Y_0 follow the assumed $N(X\beta,\sigma^2 I)$ distribution of model (2.2) while the last n_0-n elements are degenerate at (Z,U). The influence function (2.8) then explicitly measures the effect of adding the n_0-n observations (Z,U) to the original model:

$$I_\beta(F,G,\psi) = (X'\dot{D}_\psi X)^{-1}U'\Psi(Z-UT_\beta(F,\psi)) \tag{2.12}$$

and if $\psi(t_i) = t_i$,

$$I_\beta(F,G,\psi_{LS}) = (X'X)^{-1}U'(Z-U\beta). \tag{2.13}$$

By letting $X' = (X_0':U')$, $Y' = (Y_0':Z')$, and reversing the assumptions on $F(\cdot)$ and $G(\cdot)$, the effect of deleting n_0-n observations from an estimator can be determined:

$$I_\beta(F,G,\psi) = -(X'\dot{D}_\psi X)^{-1}U'\Psi(Z-UT_\beta(F,\psi)) \tag{2.14}$$

and

$$I_\beta(F,G,\psi_{LS}) = -(X'X)^{-1}U'(Z-UT_\beta(F,\psi_{LS})). \tag{2.15}$$

It is important to note that $T_\beta(F,\psi_{LS}) \neq \beta$ for this model since $F_X(Y)$ contains a $N(X_0\beta,\sigma^2 I)$ component and a point-mass at (Z,U); consequently, $T_\beta(F,\psi_{LS})$ is given by equation (2.3) with $X_+' = (X_0':U')$ and $X = X_0$.

Influence functions are valuable for investigating the effects of individual data values (y,x') on estimators as well as for studying the sensitivity of estimators to model assumptions. Each of the above influence functions appears to be an unbounded function of the elements of X_0 and the least squares influence functions are unbounded in the elements of Y_0. Most of the robust regression estimators currently in use are specifically designed to bound residuals and do provide protection against extreme response variable values (see, for example, Denby and Larson (1977)). Few provide specific protection against extreme predictor variable values.

Dorsett and Gunst (1982) argue that robust regression estimators which only weight residuals do not adequately protect against estimator distortion when extreme predictor-variable values occur in a data set (see also Mallows 1975, Denby and Larson 1977). They show that least squares, M-estimators, and Krasker and Welsch's bounded-influence estimator can be distorted when predictor-variable values are sufficiently extreme. Dorsett and Gunst propose the weighting of predictor-variable values in order to achieve a bound on the leverage values and then using M-estimation on the residuals to protect against extreme response values (see Huber 1981, p. 193, for a similar suggestion). Consideration of equations (2.13) and (2.15) for single extreme observations (z,u') leads to an alternative proposal.

2.2 Individual Extreme Observations

If a single observation at (z,u') is added to n observations which are true to the model (2.2), equation (2.3) defines the corresponding model parameter for least squares estimation. A direct measure of the change in the estimator (parameter) is $(X_+' = (X':u))$

$$T_\beta(F,\psi_{LS})-\beta = (X_+'X_+)^{-1}u(z-u'\beta)$$

$$= (X_+'X_+)^{-1}u\varepsilon_{(z)} , \qquad (2.16)$$

where $\varepsilon_{(z)} = \varepsilon_z/(1-h_u)$, $\varepsilon_z = z-u'T_\beta(F,\psi_{LS})$, and $h_u = u'(X_+'X_+)^{-1}u$. In equation (2.16) $\varepsilon_{(z)}$ is in the form of a "deleted residual" (e.g., Gunst and Mason 1980, Chapter 7). The importance of equation (2.16) is that it reveals that a single errant observation (z,u') can have a dramatic effect on least squares estimators when it is included in a data set. Another important measure of this effect is the influence function.

With obvious modifications in notation, equation (2.13) expresses the average effect on least squares estimators when an observation (z,u') is added to the model (2.2) a very small percentage of time:

$$\begin{aligned} I_\beta(F,G,\psi_{LS}) &= (X'X)^{-1}u(z-u'\beta) \\ &= (X_+'X_+)^{-1}u\varepsilon_{(z)} \quad , \end{aligned} \tag{2.17}$$

where $\varepsilon_{(z)} = \varepsilon_z/(1-h_u)$ and $\varepsilon_z = z-u'\beta = z-u'T_\beta(F,\psi_{LS})$. Interestingly, if one considers the effect of deleting an observation (z,u') a very small percentage of time from a model in which n observations are consistent with model (2.2) and the (n+1)st observation is a point mass at (z,u'), equation (2.15) defines the effect on the least squares estimator:

$$\begin{aligned} I_\beta(F,G,\psi_{LS}) &= -(X'X)^{-1}u(z-u'T_\beta(F,\psi_{LS})) \\ &= -(X'X)^{-1}u(1-h_u)(z-u'\beta) \ . \end{aligned} \tag{2.18}$$

Thus, equation (2.17) demonstrates that the infrequent inclusion of an extreme observation, particularly one with extreme predictor variable values (i.e., $h_u \approx 1$), can have a dramatic effect on the least squares estimator; on the other hand, exclusion of such an observation a small percentage of time could only have a small effect on the distorted estimator $T_\beta(F,\psi_{LS})$. One should realize that equations (2.17) and (2.18) are average effects on an estimator while equation (2.19) measures the actual effect when such an observation is included (excluded).

An important interpretive feature of equations (2.16) and (2.17) is that each can be decomposed into two components, $(X_+'X_+)^{-1}u$ and $\varepsilon_{(z)}$. The first component is a bounded function of the individual elements of u. Dorsett and Gunst (1982) demonstrate that boundedness of the influence function is not a sufficient condition for protecting robust estimators from distortion by extreme observations; however, boundedness is clearly necessary for robustness. The deleted residual $\varepsilon_{(z)}$ is an unbounded function of the individual elements of (z,u'). For fixed z and β, $I_\beta(F,G,\psi_{LS})$ is an increasing function of the leverage value h_u; consequently, $\varepsilon_{(z)}$ can be expected to be extremely large when h_u is close to one. These properties suggest that M-estimators based on deleted residuals might afford better protection from both extreme response and extreme predictor-variable values than least squares estimators and M-estimators based on raw residuals. Further support for this proposal is obtained from sample-based influence functions.

3. SAMPLE-BASED INFLUENCE FUNCTIONS

Two of the more frequently advocated types of sample-based influence functions are termed the "empirical influence function" and the "sample influence function" (Mallows 1975, Cook and

Weisberg 1982). Let $T_\beta(F_n,\psi) = \hat{\beta}_\psi$ denote the solution to

$$X'\Psi(Y-X\hat{\beta}_\psi) = 0 \; ; \tag{3.1}$$

i.e., $T_\beta(F_n,\psi)$ is a finite-sample estimator of $T_\beta(F,\psi)$. An empirical influence function replaces the functional (parameter, large-sample estimator) $T_\beta(F,\psi)$ by $T_\beta(F_n,\psi)$ in a theoretical expression for an influence function. For example, if $T_\beta(F_n,\psi_{LS}) = \hat{\beta}$ denotes the least squares estimator obtained from equation (3.1) by letting $\psi(t_i) = t_i$, the empirical influence function corresponding to equation (2.11) is

$$EIF_\beta(F_n,\psi_{LS};z,u') = n(X'X)^{-1}u(z-u'\hat{\beta}). \tag{3.2}$$

Similarly, the empirical influence corresponding to equation (2.9) is

$$EIF_\beta(F_n,\psi;Y_0,X_0) = (X'X)^{-1}X_0'\Psi(Y_0-X_0\hat{\beta}_\psi), \tag{3.3}$$

where $\hat{\beta}_\psi$ is the solution to equation (3.1). Each of these empirical influence functions are valuable for studying the effects of arbitrary observations (z,u') or (Y_0,X_0) on an estimator. Each of these functions can also be evaluated at specific data values (y_i,x_i') but other sample-based influence functions are more appropriate for this purpose.

The sample influence function measures the change in coefficient estimates when (y_i,x_i') is eliminated from the data set:

$$SIF_\beta(F_n,\psi;y_i,x_i') = T_\beta(F_n,\psi) - T_\beta(F_{(i)},\psi) \; , \tag{3.4}$$

where $T_\beta(F_{(i)},\psi)$ denotes the solution to equation (3.1) when (y_i,x_i') is eliminated from the data set. Although for arbitrary $\psi(\cdot)$ the sample influence function has no closed-form algebraic representation, it can be numerically evaluated once the respective estimates are obtained. For least squares estimation there is a closed-form algebraic representation:

$$SIF_{\beta}(F_n,\psi_{LS};y_i,x_i') = (X'X)^{-1}x_i r_{(i)} \quad , \qquad (3.5)$$

where $r_{(i)} = (y_i - x_i'\hat{\beta})/(1-h_i)$ and $h_i = x_i'(X'X)^{-1}x_i$. When $(z,u') = (y_i,x_i')$, this sample influence function is comparable with its theoretical equivalent, equation (2.16).

One advantage to the definition of the general influence function (2.8) is that many of the alternative sample-based influence functions are special cases of empirical influence functions derived from this theoretical influence function. The empirical influence functions (3.2) and (3.3) are obviously special cases. The sample influence function is an empirical influence function corresponding to equation (2.17) with $(z,u') = (y_i,x_i')$. Thus there is no need to separately define sample influence functions or to use alternative terminology; e.g., equation (3.5) is an empirical influence function, sample influence function, and is called $DFBETA_i$ by Belsley, Kuh and Welsch (1980). In addition, many other sample-based influence functions can be expressed as empirical influence functions which are special cases of the general influence function (2.8); e.g., Tukey's sensitivity curves (Huber, 1981).

Empirical influence functions are multivariate and often difficult to use in practice. Univariate statistics based on these influence functions have been proposed for use in evaluating the effects of individual observations on least squares estimates. Two of the more popular statistics are studentized (deleted) residuals (e.g., Gunst and Mason 1980)

$$t_{(i)} = r_{(i)}/\{\hat{\sigma}^2_{(i)}/(1-h_i)\}^{\frac{1}{2}}$$
$$= r_i/\{\hat{\sigma}_{(i)}(1-h_i)^{\frac{1}{2}}\} \quad (3.6)$$

and Cook's D_i (Cook 1977)

$$D_i = (\beta-\hat{\beta}_{(i)})'X'X(\beta-\hat{\beta}_{(i)})/m\hat{\sigma}^2$$
$$= h_i r^2_{(i)}/m\hat{\sigma}^2 \quad (3.7)$$

It is interesting to observe that in each of equations (3.5), (3.6) and (3.7) the deleted residual $r_{(i)}$ is a major component of the empirical influence function. This fact reinforces the important role of the deleted residual in measures of influence, which was noted in Section 2. It again suggests that M-estimators which bound the magnitude of deleted residuals might afford better protection from both extreme response and extreme predictor variable values than least squares estimators or M-estimators which bound the magnitude of raw residuals.

4. ESTIMATION WITH DELETED RESIDUALS

In this section we investigate the performance of several M-estimators which bound the magnitudes of deleted residuals or functions of the deleted residuals. All the M-estimators except those based on studentized (deleted) residuals are defined as solutions to the following system of equations, the last one included to provide a robust estimate of the scale parameter σ:

$$\sum_{i=1}^{n} x_{ij}\psi(r_i/\hat{\sigma}_\psi) = 0 \qquad j=1,2,\ldots,m \tag{4.1}$$

and

$$\sum_{i=1}^{n} \chi(r_i/\hat{\sigma}_\psi) = (n-m)a \tag{4.2}$$

where $r_i = y_i - x_i'\hat{\beta}_\psi$, $\chi(t_i) = t_i\psi(t_i) - \rho(t_i)$, $\psi(t_i) = d\rho(t_i)/dt_i$, and $a = E[\chi(\varepsilon/\sigma)]$ with $\varepsilon \sim N(0,\sigma^2)$. Huber and Dutter's H-algorithm (Dutter 1977, Huber 1981) is used to solve (4.1) and (4.2). Two choices for $\psi(\cdot)$ are studied, Huber's

$$\psi_1(t) = \begin{cases} t & |t| \leq c \\ c & |t| > c \end{cases} \tag{4.3}$$

and Hampel's redescending $\psi(\cdot)$-function

$$\psi_2(t) = \begin{cases} t & |t| \leq c_1 \\ c_1 \text{ sign}(t) & c_1 < |t| \leq c_2 \\ c_1[c_3 \text{ sign}(t) - t]/(c_3 - c_2) & c_2 < |t| \leq c_3 \\ 0 & c_3 < |t| \end{cases} \tag{4.4}$$

By choosing appropriate values for the turning constants in equations (4.3) and (4.4) the solutions to equations (4.1) and (4.2) can be made dependent on either the raw or the deleted residuals.

In choosing the values for c_2 and c_3 with Hampel's $\psi_2(\cdot)$, care must be exercized since values of c_2 and c_3 which are too close to one another can result in inflated asymptotic variances for $\hat{\beta}_\psi$ (see Huber 1981, p. 103). Observations with $c_2 < |r_i/\hat{\sigma}_\psi| \leq c_3$ can also affect scale estimates in the iteration process when Hampel's $\psi_2(\cdot)$ is used. For these reasons, when Hampel's $\psi_2(\cdot)$ is used in conjunction with either raw or deleted residuals we will estimate σ using equations (4.1) and (4.2) with Huber's $\psi_1(\cdot)$ and then insert this estimate of σ into equations (4.1) with Hampel's $\psi_2(\cdot)$ to estimate β. Once a solution to equations (4.1) with Hampel's $\psi_2(\cdot)$ is attained, a final estimate of σ is obtained from equation (4.2):

$$\hat{\sigma}_\psi^2 = \sum_{i=1}^{n} \chi(r_1/\tilde{\sigma}_\psi)\tilde{\sigma}_\psi^2/(n-m)a,$$

where $\tilde{\sigma}_\psi$ is the estimate of σ calculated with Huber's $\psi_1(\cdot)$.

Several M-estimators will be compared with the least squares estimator $\hat{\beta}$. The first two utilize Huber and Hampel's $\psi(\cdot)$-functions on raw residuals r_i and will be denoted $\hat{\beta}_1(r_i)$ and $\hat{\beta}_2(r_i)$, respectively. The turning constant for Huber's $\psi_1(\cdot)$-function is 1.345, which yields 95% efficiency for location estimation at the normal model. The turning constants for Hampel's $\psi_2(\cdot)$-function are $c_1 = 1.7$, $c_2 = 3.4$, and $c_3 = 8.5$, values which were found effective by Andrews et al. (1972) for robust location estimates. Two additional estimators, $\hat{\beta}_1(r_{(i)})$ and $\hat{\beta}_2(r_{(i)})$, bound the deleted residuals $r_{(i)}$ rather than the raw residuals with the above $\psi(\cdot)$ functions. This can be accomplished by multiplying each of the turning constants in equations (4.3) and (4.4) by $(1-h_i)$, yielding different turning constants for each residual r_i.

Since studentized (deleted) residuals (equation (3.6)) have proven effective in isolating unusual response variable values, two estimators are defined using only the estimating equations (4.1) and $\psi(\cdot)$-weights based on the $t_{(i)}$. These two estimators, denoted $\hat{\beta}_1(t_{(i)})$ and $\hat{\beta}_2(t_{(i)})$, again utilize Huber's and Hampel's $\psi(\cdot)$-functions but with different turning constants. Individually the $t_{(i)}$ are Student-t variates with (n-m-1) degrees of freedom. Consequently, we feel that it is desirable to increase Huber's turning constant to 1.5 and will use $c_1 = 1.5$, $c_2 = t_{.05}(n-m-1)$, and $c_3 = t_{.01}(n-m-1)$ with Hampel's $\psi_2(\cdot)$, where $t_\alpha(\nu)$ is the 100α% (two-tailed) percentage point of a Student-t distribution with ν degrees of freedom. By multiplying the respective turning constants by $\hat{\sigma}_{(i)}(1-h_i)^{\frac{1}{2}}$, equations (4.1) and (4.3) or (4.4) can be used to solve for these estimates. The estimator $\hat{\beta}_1(t_i)$ is similar to one proposed by Hill (1977) and Welsch (1977).

Unlike the calculation of least squares deleted residuals, the computation of deleted residuals for M-estimators can be a formidable task. Least squares deleted residuals (and studentized (deleted) residuals) can be calculated from the raw residuals and leverage values from one fit to the full data set: e.g., $r_{(i)} = r_i/(1-h_i)$. M-estimator deleted residuals must be calculated from n separate estimates $\hat{\beta}_{\psi(i)}$ (in which (y_i, x_i') is deleted from the data set) at each iteration of equations (4.1), each estimate $\hat{\beta}_{\psi(i)}$ requiring separate iterations. In the appendix we outline two approximations to $\hat{\beta}_{\psi(i)}$ which can be obtained from only one set of iterations using equations (4.1) and the full data set. The approximation used in the following analyses is similar to that for the least squares $t_{(i)}$ in equation (3.6).

The final estimator to be compared is the median-centered bounded-leverage estimator which was found to be effective against extreme response and predictor-variable values in Dorsett and Gunst (1982). This estimator centers the nonconstant predictor variables by subtracting the medians, M_j, from the n observations on each variable. Then weights w_i are found so the leverage values of the matrix

$$H_W = W^{\frac{1}{2}}X(X'WX)^{-1}X'W^{\frac{1}{2}}$$

are bounded by $\eta^2 = (2m-1)/n$, where $X = [x_{ij} - M_j]$. The weights w_i are applied to the centered predictor-variable values to yield weighted observations $\tilde{x}_{ij} = M_j + w_i(x_{ij} - M_j)$ and the weighted observations are used in place of the raw predictor-variable values in equations (4.1) and (4.2). Bounding the leverage values of the weighted observations insures that extreme predictor-variable values cannot severely distort the M-estimates whereas M-estimation protects against extreme response values. This weighting scheme also bounds the influence function with respect to individual response and predictor variable values. We denote these estimators by $\tilde{\beta}_1(\cdot)$ and $\tilde{\beta}_2(\cdot)$.

4.1 Mickey-Dunn-Clark Data

This two-variable data set (Mickey, Dunn and Clark 1967) has been used by several authors to illustrate outlier detection or robust regression techniques (e.g., Andrews and Pregibon 1978, Dempster and Gasko-Green 1981, Draper and John 1981, Dorsett and Gunst 1982). In this data set observation 19 has an unusually large response value (Gesell Adaptive Score) and observation 18 has an unusually large predictor-variable value (Age in months), although the observation (y_{18},x_{18}) is consistent with the overall decreasing linear trend in the data. In order to assess the effectiveness of several robust regression estimators in compensating for extreme predictor variable values which are not consistent with the overall trend in the data, Dorsett and Gunst (1982) change one of the observations $(y_{10},x_{10}) = (94,20)$ to $(94,50)$. A second observation (y_6,x_6) is also changed from $(87,20)$ to $(87,50)$ in order to investigate the problem of "masking" by extreme observations which appear in the same region of the space of predictor variables. Tables 1 through 4 display results of an analysis of these data using the various M-estimators described above.

In Table 1 both least squares and M-estimators based on raw residuals are distorted by changing x_6 and/or x_{10} from 20 to 50. One can show that as $x_i \to \infty$ the slope estimates approach zero and the intercept estimates approach $\bar{y}_{(i)}$ for these estimators. M-estimators based on deleted residuals or studentized deleted residuals are effective in protecting against a single extreme observation but they do not adequately compensate for masked points. In spite of this last remark, the M-estimators based on deleted residuals are more effective than those based on raw residuals when isolated extreme values occur.

[Insert Table 1]

The M-estimator weights $\phi(t) = \psi(t)/t$ exhibited in Table 2 help explain the performance of these estimators. With the original data, all the M-estimators heavily weight the observation which has an extreme response value, observation 19; in fact, $\hat{\beta}_2(t_{(i)})$

[Insert Table 2]

assigns it a weight of zero. When x_{10} is changed to 50, all the M-estimator weights for this observation are larger than in the original data set, more so for the estimators based on raw residuals than the other estimators. In addition, the weight on (y_{10}, x_{10}) for $\hat{\beta}_1(r_i)$ is much less than that for the other two estimators using Huber's $\psi_1(\cdot)$ and $\hat{\beta}_2(r_i)$ does not weight this extreme point. This explains the distortion of $\hat{\beta}_1(r_i)$ and $\hat{\beta}_2(r_i)$ in Table 1 for $x_{10} = 50$. Table 2(c) shows that none of the estimators weight observations 6 or 10 when $x_6 = x_{10} = 50$ and that the weights assigned to observation 19 are larger than those for the original data set. The reinforcement of x_{10} by x_6 causes the leverage value for observation 10 to drop from 0.523 to 0.349 and the studentized (deleted) residual to drop from 2.585 to 1.498, thus resulting in the increased weights for these points and the consequent distortion of all the M-estimators based on deleted and studentized (deleted) residuals.

Coefficient estimates for the bounded-leverage estimators are shown in Table 3. These estimators, whether based on raw, deleted, or studentized (deleted) residuals, yield estimates which are similar to one another and to the least squares and M-estimators for the original data set. In addition, each estimator produces identical estimates as x_6 and x_{10} are changed. In the second column of Table 4 the weights assigned to the predictor variable values are displayed. Observations 18, which has the most extreme predictor-variable value in the original data set, is given a

[Insert Tables 3 and 4]

small weight, as are x_6 and x_{10} when they become extreme. The bounded-leverage weights for the predictor variable values produce $\tilde{x}_i$ values of 17.308 for x_6, x_{10}, and x_{18} in all three data sets. Thus, the bounded-leverage estimator weights both individual and masked predictor-variable values. The identical weighted values explain the identical M-estimator results in each row of Table 3 and in the last six columns of Table 4(a), (b), and (c).

After weighting the predictor variables, the M-estimator weights for observations 18 and 19 in Table 4 occur because each response value is now extreme for its predictor-variable value in the overall trend of the data. The M-estimator weights for $\tilde{\beta}_1(r_i)$ and $\tilde{\beta}_1(r_{(i)})$ are simialr to one another and slightly smaller than those for $\tilde{\beta}_1(t_{(i)})$. The weights for Hampel's $\psi_2(\cdot)$ are larger than those for Huber's $\psi_1(\cdot)$.

Convergence problems occured with $\tilde{\beta}_2(t_{(i)})$. After 200 iterations the slope estimate was close to zero and convergence had not been achieved. The weights $\phi_2(t_{(i)})$ revealed that observation 18 was given a weight of zero. Since observation 18 was assigned a weight of zero, it was deleted from the data set and the estimates were recomputed. The subsequent estimates are reported in Table 3 and are consistent with the other estimates reported; however, observation 19 is assigned a weight of zero in this second analysis. The bounded-leverage estimator $\tilde{\beta}_2(t_{(i)})$ effectively eliminates observations 18 and 19 from the data set.

Only three of the estimators investigated in this section are included in the analysis presented in the next section. Because of the theoretical and computational difficulties associated with Hampel's $\psi_2(\cdot)$ which were mentioned earlier and since none of these estimators appeared to be superior to those computed from Huber's $\psi_1(\cdot)$ in Tables 1 to 4, none of the estimators using Hampel's $\psi_2(\cdot)$ will be discussed further. In addition, the estimators based on studentized (deleted) residuals do not outperform those based on deleted residuals in Tables 1 to 4. The lack of proofs for convergence of estimators based on the $t_{(i)}$ and the difficult computations needed to obtain M-estimators using the $t_{(i)}$ also lead one to prefer estimators based on raw or deleted residuals. For those reasons we examine only $\hat{\beta}_1(r_{(i)})$, $\tilde{\beta}_1(r_i)$ and $\tilde{\tilde{\beta}}_1(r_{(i)})$ in the next section.

4.2 Lave and Seskin Data

Lave and Seskin (1970, 1977, 1979) compile a large data set as part of their study of the effects of air pollution on human health. Gibbons and McDonald (1980a, 1980b, 1982) discuss several aspects of the analysis of this data set including the choice of a suitable mortality index, selection of predictor variables, and identification of influential observations. Tables 5 and 6 display analyses of the 1960 cross-sectional data on 117 Standard Metropolitan Statistical Areas (SMSAs) using the eleven predictor variables studied by Gibbons and McDonald (1982).

All three robust estimates in Table 5 for SMIN and PMEAN are smaller than those for least squares and the robust estimates for

SMEAN are larger than that for least squares. Especially large differences among the robust estimates occur for PM2, a population density variable and LN(POP), the logarithm of population size. The summary statistics in Table 6 help explain the differences in these estimates.

All the robust estimates weight several of the SMSAs very heavily. Included in the observations which are heavily weighted are Tampa, Scranton, Wilkes Barre, and Austin, all of which receive a weight less than 0.5 by all the robust estimates. These four SMSAs have the largest studentized (deleted) residuals, indicating (since none have extremely large leverage values) that their response values are extreme relative to their predictor-variable values.

The leverage values in Table 6 indicate that Jersey City is an extreme point in the space of predictor variables. This is primarily due to its extremely large value for PM2. The bounded-leverage estimators weight Jersey City's predictor variables and then the M-estimator weights the the resulting residual. Note that the M-estimator using deleted residuals does not weight Jersey City. Gibbons and McDonald (1982) discovered that Jersey City has a great influence on the least squares coefficient estimate for PM2 by calculating Belsley, Kuh, and Welsch's (1980) DFBETAS. Thus the difference between the least squares and M-estimator estimates for PM2 and the bounded-leverage estimates for PM2 is primarily due to the weighting of Jersey City.

Charleston and Tampa both influence the least squares estimate for LN(POP) (Gibbons and McDonald 1982). All three robust estimators in Table 5 weight the residuals for these two SMSAs but the bounded-leverage estimators also weight the predictor-variable values. The bounded-leverage estimates also weight the value of LN(POP) for New York, the SMSA which has the most extreme value for LN(POP). This difference in weighting accounts for the difference in the estimates for $\hat{\beta}_1(r_{(i)})$ and those for the bounded-leverage estimators.

Overall, the bounded-leverage estimates appear to be preferable to the M-esimator $\hat{\beta}_1(r_{(i)})$ for this data set. The bounded-leverage estimates are weighting influential observations regardless of whether the influential observation is due to extreme predictor values or extreme response values. In particular, the weighting of Jersey City and New York seems warranted. Although the two bounded-leverage estimators produce similar estimates for this data set, the added protection afforded by M-estimation on deleted residuals suggests that $\tilde{\beta}_1(r_{(i)})$ is to be preferred to $\tilde{\beta}_1(r_i)$. This recommendation is reinforced by the expressions for regression influence functions which were derived in Sections 2 and 3.

5. CONCLUSION

Theoretical and sample-based influence functions indicate the important role of deleted residuals in measuring the effect of individual observations on regression estimators, notably least squares estimators. Although bounding the magnitude of deleted residuals improves the performance of regression M-estimators

APPENDIX

Studentized (deleted) residuals for M-estimates have the form

$$t_{(i)} = (y_i - x_i'\hat{\beta}_{\psi(i)})/\{\hat{\sigma}_{\psi(i)}(1-h_{\psi(i)})^{\frac{1}{2}}\} \; ,$$

where $\hat{\beta}_{\psi(i)}$ is an M-estimate calculated from equations (4.1) from the reduced data set in which (y_i, x_i') has been eliminated. Calculation of $\hat{\beta}_1(t_{(i)})$ and $\hat{\beta}_2(t_{(i)})$ requires (a) an iterative solution of equations (4.1) with n new values of $t_{(i)}$ at each iteration and (b) iterative estimation of $\hat{\beta}_{\psi(i)}$ for i=1,2,...,n to obtain the $t_{(i)}$. Two approximations to the $t_{(i)}$ were investigated, each of which only uses the current estimate $\hat{\beta}_1(t_{(i)})$ or $\hat{\beta}_2(t_{(i)})$ in place of $\hat{\beta}_{\psi(i)}$ in the above equation for $t_{(i)}$.

Let $\Phi_k = \text{diag}(\phi_1^k, \ldots, \phi_n^k)$, where $\phi_i^k = \psi_k(t_{(i)})/t_{(i)}$. Solutions to equations (4.1) can be obtained through iteratively reweighted least squares by use of the following iteration formula:

$$\hat{\beta}_k(t_{(i)}) = (X'\Phi_k X)^{-1} X'\Phi_k Y \quad , \quad k=1,2.$$

The first approximation to the $t_{(i)}$ uses $\hat{\beta}_k(t_{(i)})$ in place of $\hat{\beta}_{\psi(i)}$,

$$\hat{\sigma}^2_{\psi(i)} = \{\sum_{j=1}^{n} \tilde{r}_j^2 - (1-h_{\psi(i)})^{-2} r_i^2\}/(n-m-1) \quad ,$$

where $h_{\psi(i)}$ is the ith diagonal element of $H_\psi = \Phi_k^{\frac{1}{2}} X(X'\Phi_k X)^{-1} X'\Phi_k^{\frac{1}{2}}$, $r_i = y_i - x_i'\hat{\beta}_k(t_{(i)})$, and

$$\tilde{r}_j = r_j + (1-h_{\psi(i)})^{-1} \phi_i^k r_i u_j' (X'\Phi_k X)^{-1} u_i .$$

The adjustments in $\tilde{r}_j$ and $\hat{\sigma}^2_{\psi(i)}$ are intended to accomodate for the use of $\hat{\beta}_k(t_{(i)})$ rather than the individual $\hat{\beta}_{\psi(i)}$ (see Henderson (1982)). In several test cases the estimates produced by this

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., and TUKEY, J. W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press.

ANDREWS, D. F. and PREGIBON, D. (1978), "Finding the Outliers that Matter," *Journal of the Royal Statistical Society, Series B*, 40, 85-93.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980), *Regression Diagnostics*, New York: John Wiley and Sons, Inc.

COOK, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15-18.

COOK, R. D. and WEISBERG, S. (1980). "Characteristics of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495-508.

——— (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

DEMPSTER, A. P. and GASKO-GREEN, M. (1981). "New Tools for Residual Analysis," *The Annals of Statistics*, 9, 945-959.

DENBY, L. and LARSON, W. A. (1977), "Robust Regression Estimators Compared Via Monte Carlo," *Communications in Statistics*, A6 355-362.

DORSETT, D. and GUNST, R. F. (1982). "Bounded-Leverage Weights for Robust Regression Estimators," submitted for publication.

DRAPER, N. R. and JOHN, J. A. (1981), "Influential Observations and Outliers in Regression," *Technometrics*, 23, 21-26.

DUTTER, R. (1977), "Numerical Solution of Robust Regression Problems: Computational Aspects, a Comparison," *Journal of Statistical Computation and Simulation*, 5, 207-238.

GIBBONS, D. I. and MCDONALD, G. C. (1980a), "Sensitivity of an Air Pollution and Health Study to the Choice of a Mortality Index," General Motors Research Publication GMR-3256, General Motors Research Laboratories.

——— (1980b), "Examining Regression Relationships Between Air Pollution and Mortality," General Motors Research Publication GMR-3278, General Motors Research Laboratories.

Table 1. Comparison of Coefficient Estimates, Mickey-Dunn-Clark Data

	Original Data	$x_{10} = 50$	$x_6 = x_{10} = 50$
		(a) Intercept	
Least Squares	109.87	102.76	101.47
Huber's $\psi_1(\cdot)$			
$\hat{\beta}_1(r_i)$	109.74	102.84	100.26
$\hat{\beta}_1(r_{(i)})$	109.80	107.03	99.86
$\hat{\beta}_1(t_{(i)})$	109.58	106.32	100.37
Hampel's $\psi_2(\cdot)$			
$\hat{\beta}_2(r_i)$	109.63	102.53	100.93
$\hat{\beta}_2(r_{(i)})$	109.62	109.60	100.39
$\hat{\beta}_2(t_{(i)})$	109.31	109.84	98.88
		(b) Slope	
Least Squares	-1.13	-0.58	-0.45
Huber's $\psi_1(\cdot)$			
$\hat{\beta}_1(r_i)$	-1.17	-0.62	-0.38
$\hat{\beta}_1(r_{(i)})$	-1.17	-0.98	-0.33
$\hat{\beta}_1(t_{(i)})$	-1.16	-0.90	-0.38
Hampel's $\psi_2(\cdot)$			
$\hat{\beta}_2(r_i)$	-1.16	-0.58	-0.42
$\hat{\beta}_2(r_{(i)})$	-1.16	-1.18	-0.38
$\hat{\beta}_2(t_{(i)})$	-1.19	-1.18	-0.22

Table 2. M-Estimator Weights for Selected Points, Mickey-Dunn-Clark Data[a]

	Huber's $\psi_1(\cdot)$			Hampel's $\psi_2(\cdot)$		
Obsn.	$\phi_1(r_i)$	$\phi_1(r_{(i)})$	$\phi_1(t_{(i)})$	$\phi_2(r_i)$	$\phi_2(r_{(i)})$	$\phi_2(t_{(i)})$
			(a) Original Data			
6						
10						
18						
19	.457	.478	.472	.555	.575	.000
			(b) $x_{10} = 50$			
6						
10	.800	.436	.476		.388	.000
18	.898					
19	.619	.511	.535	.792	.597	.150
			(c) $x_6 = x_{10} = 50$			
6						
10						
18	.629	.595	.655	.850	.787	.299
19	.637	.639	.666	.812	.800	.752

[a]Only weights less than 0.9 are displayed.

Table 3. Comparison of Bounded-Leverage Estimates, Mickey-Dunn-Clark Data

	Original Data	$x_{10} = 50$	$x_6 = x_{10} = 50$
		(a) Intercept	
Least Squares	109.87	102.76	101.47
Huber's $\psi_1(\cdot)$			
$\tilde{\beta}_1(r_i)$	108.60	108.60	108.60
$\tilde{\beta}_1(r_{(i)})$	107.69	107.69	107.69
$\tilde{\beta}_1(t_{(i)})$	108.74	108.74	108.74
Hampel's $\psi_2(\cdot)$			
$\tilde{\beta}_2(r_i)$	108.83	108.83	108.83
$\tilde{\beta}_2(r_{(i)})$	108.50	108.50	108.50
$\tilde{\beta}_2(t_{(i)})$[a]	108.23	108.23	108.23
		(b) Slope	
Least Squares	-1.13	-0.58	-0.45
Huber's $\psi_1(\cdot)$			
$\tilde{\beta}_1(r_i)$	-1.19	-1.19	-1.19
$\tilde{\beta}_1(r_{(i)})$	-1.10	-1.10	-1.10
$\tilde{\beta}_1(t_{(i)})$	-1.20	-1.20	-1.20
Hampel's $\psi_2(\cdot)$			
$\tilde{\beta}_2(r_i)$	-1.20	-1.20	-1.20
$\tilde{\beta}_2(r_{(i)})$	-1.17	-1.17	-1.17
$\tilde{\beta}_2(t_{(i)})$[a]	-1.17	-1.17	-1.17

[a]Observation 18 deleted from the data set.

Table 4. Bounded-Leverage Weights for Selected Points, Mickey-Dunn-Clark Data[b]

		Huber's $\psi_1(\cdot)$			Hampel's $\psi_2(\cdot)$		
Obsn.	w_1	$\phi_1(r_i)$	$\phi_1(r_{(i)})$	$\phi_1(t_i)$	$\phi_2(r_i)$	$\phi_2(r_{(i)})$	$\phi_2(t_{(i)})$[a]
			(a) Original Data				
6	.799						
10	.799						
18	.232	.555	.538	.576	.688	.672	---
19		.512	.518	.530	.631	.625	.000
			(b) $x_{10} = 50$				
6	.799						
10	.184						
18	.232	.555	.538	.576	.688	.672	---
19		.512	.518	.530	.631	.625	.000
			(c) $x_6 = x_{10} = 50$				
6	.184						
10	.184						
18	.232	.555	.538	.576	.688	.672	---
19		.512	.518	.530	.631	.625	.000

[a]Observation 18 deleted from the data set.

[b]Only weights less than 0.9 are displayed.

Table 5. Comparison of Coefficient Estimates, Lave and Seskin Data.

Predictor Variable	Least Squares	Huber's $\psi_1(\cdot)$		
		$\hat{\beta}_1(r_{(i)})$	$\tilde{\beta}_1(r_i)$	$\tilde{\beta}_1(r_{(i)})$
SMIN	.456 (.287)	.244 (.239)	.274 (.260)	.230 (.256)
SMEAN	.079 (.329)	.315 (.274)	.283 (.289)	.316 (.284)
SMAX	.056 (.114)	-.009 (.095)	.006 (.103)	.008 (.101)
PMIN	.253 (.618)	.188 (.515)	.055 (.538)	.051 (.530)
PMEAN	.337 (.424)	.059 (.353)	.140 (.385)	.098 (.379)
PMAX	-.025 (.010)	.047 (.083)	.008 (.099)	.020 (.097)
PM2	.089 (.054)	.087 (.045)	.395 (.137)	.361 (.135)
GE65	6.923 (.413)	7.331 (.344)	7.179 (.373)	7.233 (.367)
PNOW	.403 (.104)	.412 (.087)	.355 (.093)	.354 (.092)
POOR	.039 (.151)	.095 (.125)	.177 (.136)	.190 (.134)
LN(POP)	-12.217 (8.72)	-9.251 (7.27)	-17.072 (8.53)	-15.988 (8.41)

Table 6. Summary Statitics for Selected Points, Lave and Seskin Data[a]

SMSA	h_i	$t_{(i)}$	Robust Estimator Weights $\hat{\phi}_1(r_{(i)})$	w_i	$\tilde{\phi}_1(r_i)$	$\tilde{\phi}_1(r_{(i)})$
Miami FL			.618		.698	.661
Orlando FL		1.847	.530		.569	.543
Tampa FL	.278	4.816	.238	.813	.461	.431
Macon GA	.422			.702		
Savannah GA		2.003	.568		.600	.577
Terre Haute IN	.102	1.854	.701		.772	.762
New Orleans LA		1.984	.574		.596	.574
Jackson MS	.100		.862			.892
Las Vegas NV			.618		.607	.578
Jersey City NJ	.894			.179	.383	.356
Albuquerque NM	.252					
New York NY	.195			.560		
Canton OH	.261	-1.982	.715	.776		
Scranton PA	.145	3.622	.326		.324	.314
Wilkes Barre PA		3.731	.324		.352	.343
Sioux Falls SD		-1.700	.597		.620	.591
Austin TX		-2.243	.486		.489	.474
Waco TX			.770		.757	.720
Charleston WV	.557	-1.694	.763	.574	.858	.765
Madison WI		-1.504	.736		.791	.766

[a]Only leverage values greater than 0.1, $|t_{(i)}|$ values greater than 1.5, and weights less than 0.9 are displayed. $\hat{\phi}_1(\cdot)$ and $\tilde{\phi}_1(\cdot)$ denote, respectively, weights for the M-estimator and the bounded-leverage estimators.