Journal Title: applied regression analysis and generalized linear models

Volume:
Issue:
Month/Year: 2008
Pages: 530-~~547~~

Article Author: John Fox

Article Title: Robust Regression

586-604

Call #: HA31.3 .F69 2008

Location: F

CUSTOMER HAS REQUESTED:
Electronic Delivery: Yes
Alternate Delivery Method:
YesHold for pickup

Jason Geller (jg9120)
101 Lassen Court Apt 9
Princeton, NJ 08904

Note

THIRD EDITION

APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS

John Fox
McMaster University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

19 Robust Regression*

The efficiency of least-squares regression is seriously impaired by heavy-tailed error distributions; in particular, least squares is vulnerable to outlying observations at high-leverage points.[1] One response to this problem is to employ diagnostics for high-leverage, influential, and outlying data; if unusual data are discovered, then these can be corrected, removed, or otherwise accommodated.

Robust estimation is an alternative approach to outliers and the heavy-tailed error distributions that tend to generate them. Properly formulated, robust estimators are almost as efficient as least squares when the error distribution is normal and much more efficient when the errors are heavy tailed. Robust estimators hold their efficiency well because they are resistant to outliers. Rather than simply discarding discrepant data, however, robust estimation (as we will see) down-weights them.

Much of the chapter is devoted to a particular strategy of robust estimation, termed *M estimation*, due originally to Huber (1964). I also describe two other approaches to robust estimation: *bounded-influence regression* and *quantile regression*. Finally, I briefly present robust estimators for generalized linear models.

19.1 *M* Estimation

19.1.1 Estimating Location

Although our proper interest is in robust estimation of linear models, it is helpful to narrow our focus initially to a simpler setting: robust estimation of *location*—that is, estimation of the center of a distribution. Let us, then, begin our exploration of robust estimation with the minimal linear model

$$Y_i = \mu + \varepsilon_i$$

where the observations Y_i are independently sampled from some symmetric distribution with center μ (and hence the errors ε_i are independently and symmetrically distributed around 0).[2]

If the distribution from which the observations are drawn is normal, then the sample mean $\widehat{\mu} = \overline{Y}$ is the maximally efficient estimator of μ, producing the fitted model

[1] See Chapter 11.

[2] In the *absence* of symmetry, what we mean by the center of the distribution becomes ambiguous.

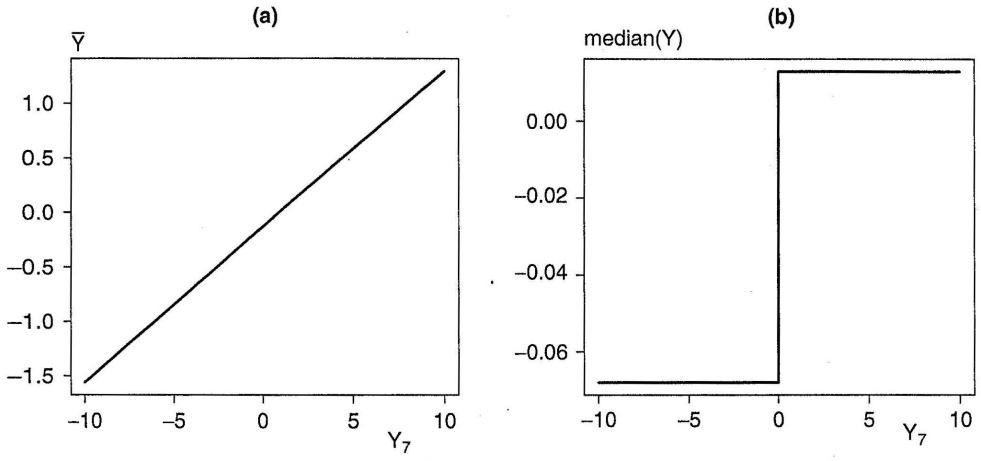

Figure 19.1 The influence functions for the mean (a) and median (b) for the sample $Y_1 = -0.068$, $Y_2 = -1.282, Y_3 = 0.013, Y_4 = 0.141, Y_5 = -0.980, Y_6 = 1.263$. The influence function for the median is bounded, while that for the mean is not. Note that the vertical axes for the two graphs have different scales.

$$Y_i = \overline{Y} + E_i$$

The mean minimizes the least-squares *objective function*:

$$\sum_{i=1}^{n} \rho_{\text{LS}}(E_i) = \sum_{i=1}^{n} \rho_{\text{LS}}(Y_i - \widehat{\mu}) \equiv \sum_{i=1}^{n} (Y_i - \widehat{\mu})^2$$

The mean, however, is very sensitive to outliers, as is simply demonstrated: I drew a sample of six observations from the standard-normal distribution, obtaining

$$Y_1 = -0.068 \quad Y_2 = -1.282 \quad Y_3 = 0.013$$
$$Y_4 = 0.141 \quad Y_5 = -0.980 \quad Y_6 = 1.263$$

The mean of these six values is $\overline{Y} = -0.152$. Now, imagine adding a seventh observation, Y_7, allowing it to take on all possible values from -10 to $+10$ (or, with greater imagination, from $-\infty$ to $+\infty$). The result, called the *influence function* of the mean, is graphed in Figure 19.1(a). It is apparent from this figure that as the discrepant seventh observation grows more extreme, the sample mean chases it.

The shape of the influence function for the mean follows from the derivative of the least-squares objective function with respect to E:

$$\psi_{\text{LS}}(E) \equiv \rho'_{\text{LS}}(E) = 2E$$

Influence, therefore, is proportional to the residual E. It is convenient to redefine the least-squares objective function as $\rho_{\text{LS}}(E) \equiv \frac{1}{2}E^2$, so that $\psi_{\text{LS}}(E) = E$.

Now consider the sample median as an estimator of μ. The median minimizes the *least-absolute-values* (LAV) objective function:[3]

[3] See Exercise 19.1.

$$\sum_{i=1}^{n} \rho_{\text{LAV}}(E_i) = \sum_{i=1}^{n} \rho_{\text{LAV}}(Y_i - \widehat{\mu}) \equiv \sum_{i=1}^{n} |Y_i - \widehat{\mu}|$$

As a result, the median is much more resistant than the mean to outliers. The influence function of the median for the illustrative sample is shown in Figure 19.1(b). In contrast to the mean, the influence of a discrepant observation on the median is *bounded*. Once again, the derivative of the objective function gives the shape of the influence function:[4]

$$\psi_{\text{LAV}}(E) \equiv \rho'_{\text{LAV}}(E) = \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}$$

Although the median is more resistant than the mean to outliers, it is less efficient than the mean if the distribution of Y is normal. When $Y \sim N(\mu, \sigma^2)$, the sampling variance of the mean is σ^2/n, while the variance of the median is $\pi\sigma^2/2n$: That is, $\pi/2 \approx 1.57$ times as large as for the mean. Other objective functions combine resistance to outliers with greater robustness of efficiency. Estimators that can be expressed as minimizing an objective function $\sum_{i=1}^{n} \rho(E)$ are called *M estimators*.[5]

Two common choices of objective functions are the *Huber* and Tukey's *biweight* (or *bisquare*) functions:

- The Huber objective function is a compromise between least squares and least absolute values, behaving like least squares in the center and like least absolute values in the tails:

$$\rho_H(E) = \begin{cases} \frac{1}{2}E^2 & \text{for } |E| \leq k \\ k|E| - \frac{1}{2}k^2 & \text{for } |E| > k \end{cases}$$

The Huber objective function ρ_H and its derivative, the influence function ψ_H, are graphed in Figure 19.2:[6]

$$\psi_H(E) = \begin{cases} k & \text{for } E > k \\ E & \text{for } |E| \leq k \\ -k & \text{for } E < -k \end{cases}$$

The value k, which defines the center and tails, is called a *tuning constant*.

It is most natural to express the tuning constant as a multiple of the *scale* (i.e., the spread) of the variable Y, that is, to take $k = cS$, where S is a measure of scale. The sample standard deviation is a poor measure of scale in this context because it is even more affected than the mean by outliers. A common robust measure of scale is the *median absolute deviation* (MAD):

$$\text{MAD} \equiv \text{median}|Y_i - \widehat{\mu}|$$

[4]Strictly speaking, the derivative of ρ_{LAV} is undefined at $E = 0$, but setting $\psi_{\text{LAV}}(0) \equiv 0$ is convenient.

[5]Estimators that can be written in this form can be thought of as generalizations of maximum-likelihood estimators, hence the term *M* estimator. The maximum-likelihood estimator is produced by taking $\rho_{\text{ML}}(y - \mu) \equiv -\log_e p(y - \mu)$ for an appropriate probability or probability density function $p(\cdot)$.

[6]My terminology here is loose but convenient: Strictly speaking, the ψ-function is not the influence function, but it has the same shape as the influence function.

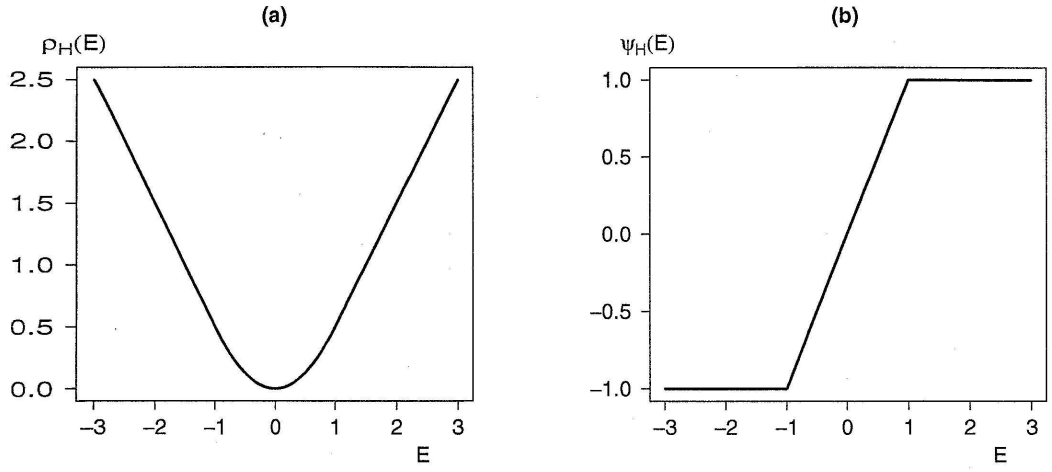

Figure 19.2 Huber objective function ρ_H (a) and "influence function" ψ_H (b). To calibrate these graphs, the tuning constant is set to $k = 1$. (See the text for a discussion of the tuning constant.)

The estimate $\widehat{\mu}$ can be taken, at least initially, as the median value of Y. We can then define $S \equiv \text{MAD}/0.6745$, which ensures that S estimates the standard deviation σ when the population is normal. Using $k = 1.345S$ (i.e., $1.345/0.6745 \approx 2$ MADs) produces 95% efficiency relative to the sample mean when the population is normal, along with considerable resistance to outliers when it is not. A smaller tuning constant can be employed for more resistance.

- The biweight (or bisquare) objective function levels off at very large residuals:[7]

$$\rho_{\text{BW}}(E) \equiv \begin{cases} \dfrac{k^2}{6}\left\{1 - \left[1 - \left(\dfrac{E}{k}\right)^2\right]^3\right\} & \text{for } |E| \le k \\ \dfrac{k^2}{6} & \text{for } |E| > k \end{cases}$$

The influence function for the biweight estimator, therefore, "redescends" to 0, *completely discounting* observations that are sufficiently discrepant:

$$\psi_{\text{BW}}(E) = \begin{cases} E\left[1 - \left(\dfrac{E}{k}\right)^2\right]^2 & \text{for } |E| \le k \\ 0 & \text{for } |E| > k \end{cases}$$

The functions ρ_{BW} and ψ_{BW} are graphed in Figure 19.3. Using $k = 4.685S$ (i.e., $4.685/0.6745 \approx 7$ MADs) produces 95% efficiency when sampling from a normal population.

[7] The term bisquare applies literally to the ψ-function and to the weight function (hence biweight) to be introduced presently—not to the objective function.

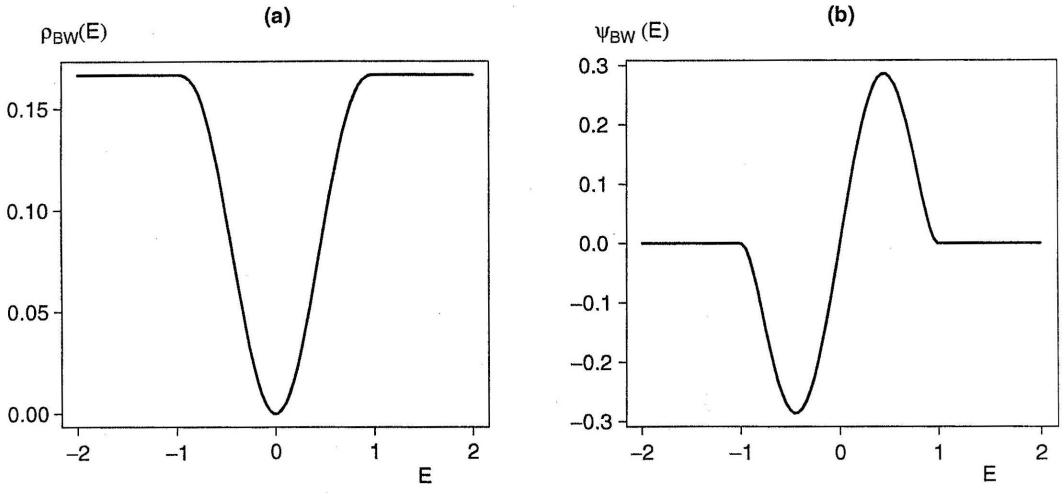

Figure 19.3 Biweight objective function ρ_{BW} (a) and "influence function" ψ_{BW} (b). To calibrate these graphs, the tuning constant is set to $k = 1$. The influence function "redescends" to 0 when $|E|$ is large.

Robust M estimators of location, for the parameter μ in the simple model $Y_i = \mu + \varepsilon_i$, minimize the objective function $\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \widehat{\mu})$, selecting $\rho(\cdot)$ so that the estimator is relatively unaffected by outlying values. Two common choices of objective function are the Huber and the biweight (or bisquare). The sensitivity of an M estimator to individual observations is expressed by the influence function of the estimator, which has the same shape as the derivative of the objective function, $\psi(E) \equiv \rho'(E)$.

Calculation of M estimators usually requires an iterative procedure (although iteration is not necessary for the mean and median, which, as we have seen, fit into the M estimation framework). An estimating equation for $\widehat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\widehat{\mu}$) to 0, obtaining

$$\sum_{i=1}^{n} \psi\left(Y_i - \widehat{\mu}\right) = 0 \tag{19.1}$$

There are several general approaches to solving Equation 19.1; probably the most straightforward, and the simplest to implement computationally, is to reweight the mean iteratively—a special case of *iterative weighted least squares* (*IWLS*):[8]

[8]In Chapter 15, we employed IWLS estimation for generalized linear models. The method is also called *iteratively reweighted least squares* (*IRLS*).

Table 19.1 Weight Functions $w(E) = \psi(E)/E$ for Several *M* Estimators

Estimator	*Weight Function w(E)*	
Least squares	1	
Least absolute values	$1/\lvert E\rvert$	(for $E \neq 0$)
Huber	1	for $\lvert E\rvert \leq k$
	$k/\lvert E\rvert$	for $\lvert E\rvert > k$
Bisquare (biweight)	$\left[1 - \left(\frac{E}{k}\right)^2\right]^2$	for $\lvert E\rvert \leq k$
	0	for $\lvert E\rvert > k$

1. Define the *weight function* $w(E) \equiv \psi(E)/E$. Then, the estimating equation becomes

$$\sum_{i=1}^{n}(Y_i - \widehat{\mu})w_i = 0 \tag{19.2}$$

where

$$w_i \equiv w(Y_i - \widehat{\mu})$$

The solution of Equation 19.2 is the weighted mean

$$\widehat{\mu} = \frac{\sum w_i Y_i}{\sum w_i}$$

The weight functions corresponding to the least-squares, LAV, Huber, and bisquare objective functions are shown in Table 19.1 and graphed in Figure 19.4. The least-squares weight function accords equal weight to each observation, while the bisquare gives 0 weight to observations that are sufficiently outlying; the LAV and Huber weight functions descend toward 0 but never quite reach it.

2. Select an initial estimate $\widehat{\mu}^{(0)}$, such as the median of the Y values.[9] Using $\widehat{\mu}^{(0)}$, calculate an initial estimate of scale $S^{(0)}$ and initial weights $w_i^{(0)} = w(Y_i - \widehat{\mu}^{(0)})$. Set the iteration counter $l = 0$. The scale is required to calculate the tuning constant $k = cS$ (for prespecified c).
3. At each iteration l, calculate $\widehat{\mu}^{(l)} = \sum w_i^{(l-1)} Y_i / \sum w_i^{(l-1)}$. Stop when the change in $\widehat{\mu}^{(l)}$ is negligible from one iteration to the next.

An estimating equation for $\widehat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\widehat{\mu}$) to 0, obtaining $\sum_{i=1}^{n} \psi(Y_i - \widehat{\mu}) = 0$. The simplest procedure for solving this estimating equation is by iteratively reweighted means. Defining the weight function as $w(E) \equiv \psi(E)/E$, the estimating equation becomes $\sum_{i=1}^{n}(Y_i - \widehat{\mu})w_i = 0$, from which $\widehat{\mu} = \sum w_i Y_i / \sum w_i$. Starting with an initial estimate $\widehat{\mu}^{(0)}$, initial weights are calculated, and the value of $\widehat{\mu}$ is updated. This procedure continues iteratively until the value of $\widehat{\mu}$ converges.

[9] Because the estimating equation for redescending *M* estimators, such as the bisquare, can have more than one root, the selection of an initial estimate might be consequential.

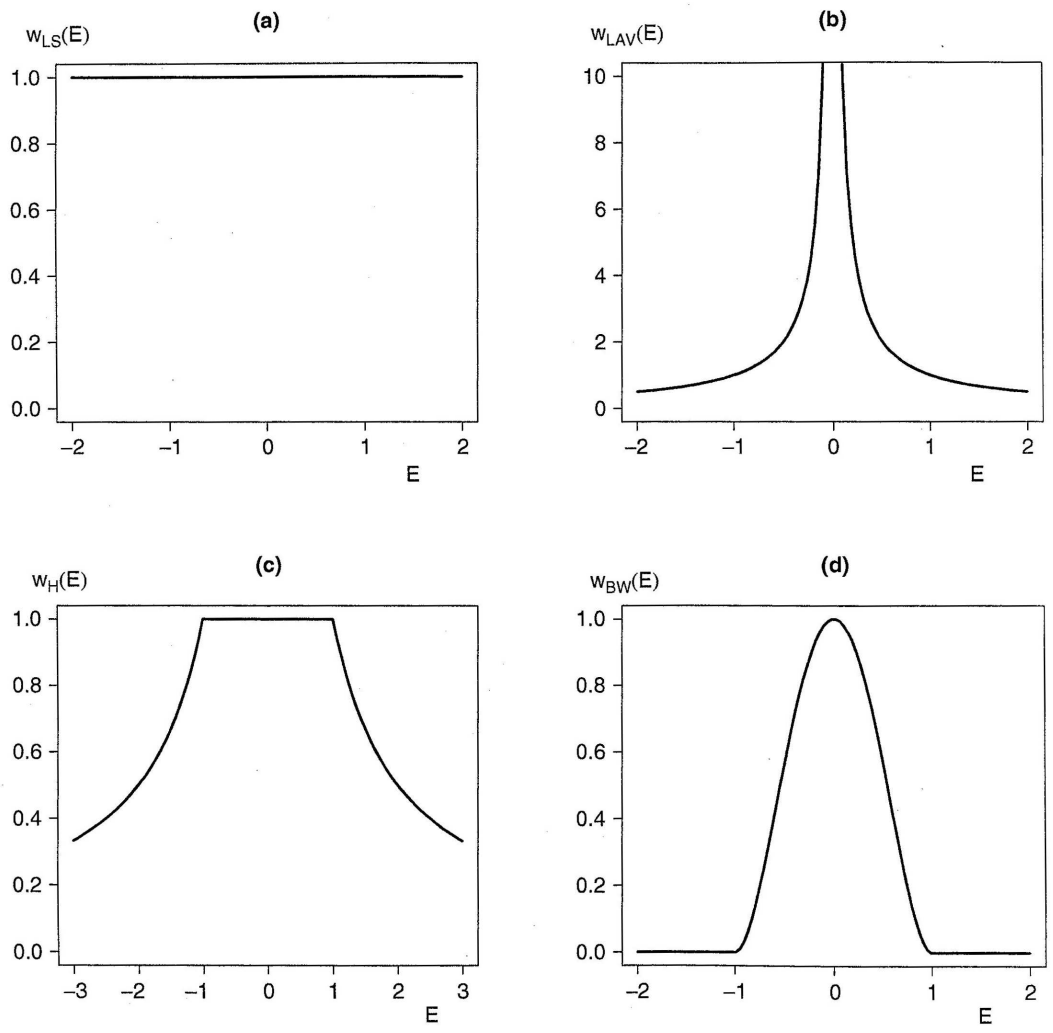

Figure 19.4 Weight functions $w(E)$ for the (a) least-squares, (b) least-absolute-values, (c) Huber, and (d) biweight estimators. The tuning constants for the Huber and biweight estimators are taken as $k = 1$. Note that the vertical axis in the graph for the LAV estimator and the horizontal axis in the graph for the Huber estimator are different from the others.

19.1.2 *M* Estimation in Regression

With the exception of one significant caveat, to be addressed in the next section, the generalization of M estimators to regression is immediate. We now wish to estimate the linear model

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \\ &= \underset{(1\times k+1)}{\mathbf{x}_i'} \quad \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \varepsilon_i \end{aligned}$$

The estimated model is

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + \cdots + B_k X_{ik} + E_i \\ &= \mathbf{x}_i'\mathbf{b} + E_i \end{aligned}$$

The general M estimator minimizes the objective function

$$\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i'\mathbf{b})$$

Differentiating the objective function and setting the derivative to $\mathbf{0}$ produces

$$\sum_{i=1}^{n} \psi(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i = \mathbf{0} \tag{19.3}$$

which is a system of $k+1$ estimating equations in the $k+1$ elements of $\mathbf{b}$.

> M estimation for the regression model $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ is a direct extension of M estimation of location: We seek to minimize an objective function of the regression residuals, $\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i'\mathbf{b})$. Differentiating the objective function and setting the derivatives to 0 produces the estimating equations $\sum_{i=1}^{n} \psi(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i = \mathbf{0}$.

Using the weight function $w(E) \equiv \psi(E)/E$ and letting $w_i \equiv w(E_i)$, the estimating equations become

$$\sum_{i=1}^{n} w_i(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i = \mathbf{0}$$

The solution to these estimating equations minimizes the weighted sum of squares $\sum w_i E_i^2$.[10] Because the weights depend on the residuals, the estimated coefficients depend on the weights, and the residuals depend on the estimated coefficients, an iterative solution is required. The IWLS algorithm for regression is as follows:

1. Select initial estimates $\mathbf{b}^{(0)}$ and set the iteration counter $l = 0$. Using the initial estimates, find residuals $E_i^{(0)} = Y_i - \mathbf{x}_i'\mathbf{b}^{(0)}$, and from these, calculate the estimated scale of the residuals $S^{(0)}$ and the weights $w_i^{(0)} = w(E_i^{(0)})$.
2. At each iteration l, solve the estimating equations using the current weights, minimizing $\sum w_i^{(l-1)} E_i^2$ to obtain $\mathbf{b}^{(l)}$. The solution is conveniently expressed as

$$\mathbf{b}^{(l)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where the model matrix $\underset{(n\times k+1)}{\mathbf{X}}$ has $\mathbf{x}_i'$ as its ith row, and $\underset{(n\times n)}{\mathbf{W}} \equiv \text{diag}\{w_i^{(l-1)}\}$. Continue until $\mathbf{b}^{(l)} - \mathbf{b}^{(l-1)} \approx \mathbf{0}$.[11]

[10] See the discussion of weighted-least-squares regression in Section 12.2.2.

[11] As in the location problem, it is possible that the estimating equations for a redescending estimator have more than one root. If you use the bisquare estimator, for example, it is prudent to pick a good start value, such as provided by the Huber estimator.

Table 19.2 *M* Estimates for Duncan's Regression of Occupational Prestige on Income and Education for 45 U.S. Occupations

	Coefficient		
Estimator	*Constant*	*Income*	*Education*
Least squares	−6.065	0.5987	0.5458
Least squares*	−6.409	0.8674	0.3322
Least absolute values	−6.408	0.7477	0.4587
Huber	−7.111	0.7014	0.4854
Bisquare (biweight)	−7.412	0.7902	0.4186

NOTE: The estimator marked "Least squares*" omits ministers and railroad conductors.

Using the weight function, the estimating equations can be written as

$$\sum_{i=1}^{n} w_i(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i' = \mathbf{0}$$

The solution of the estimating equations then follows by weighted least squares:

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is the diagonal matrix of weights. The method of iterated weighted least squares starts with initial estimates $\mathbf{b}^{(0)}$, calculates initial residuals from these estimates, and calculates initial weights from the residuals. The weights are used to update the parameter estimates, and the procedure is iterated until it converges.

The asymptotic covariance matrix of the *M* estimator is given by

$$\mathcal{V}(\mathbf{b}) = \frac{E(\psi^2)}{[E(\psi')]^2}(\mathbf{X}'\mathbf{X})^{-1}$$

Using $\sum[\psi(E_i)]^2/n$ to estimate $E(\psi^2)$ and $[\sum\psi'(E_i)/n]^2$ to estimate $[E(\psi')]^2$ produces the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$. Research suggests, however, that these sampling variances are not to be trusted unless the sample size is large.[12]

To illustrate *M* estimation, recall Duncan's regression of occupational prestige on income and education. In our previous analysis of these data, we discovered two influential observations: *ministers* and *railroad conductors*.[13] Another observation, *reporters*, has a relatively large residual but is not influential; still another observation, *railroad engineers*, is at a high-leverage point but is not discrepant. Table 19.2 summarizes the results of estimating Duncan's regression using four *M* estimators, including ordinary least squares. (The least-squares

[12]See Li (1985, pp. 300–301). For an alternative approach that may have better small-sample properties, see Street, Carroll, and Ruppert (1988). Also see the discussion of bootstrap methods in Chapter 21.

[13]See Chapter 11.

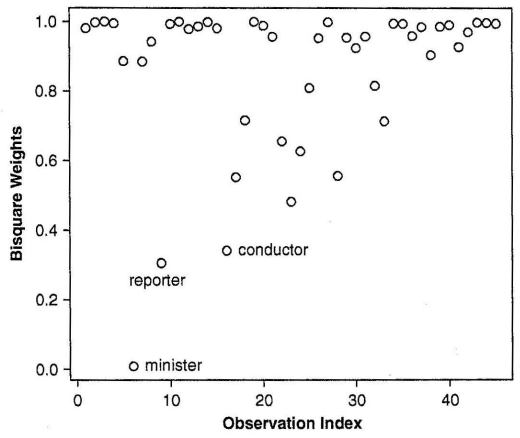

Figure 19.5 Final weights for the bisquare estimator applied to Duncan's regression of occupational prestige on income and education.

estimates obtained after deleting *ministers* and *railroad conductors* are also shown for comparison.) The three robust estimators produce quite similar results, with a larger income coefficient and smaller education coefficient than least squares. The redescending bisquare estimator is most different from least squares (and most similar to least squares after removing the two discrepant observations).

Figure 19.5 shows the final weights for the bisquare estimator applied to Duncan's data. *Railroad conductors*, *reporters*, and (especially) *ministers* have comparatively small weights, although some other occupations are down-weighted as well. Rather than simply regarding robust regression as a procedure for automatically down-weighting outliers, the method can often be used effectively, as here, to identify outlying observations.

19.2 Bounded-Influence Regression

The flies in the ointment of M estimation in regression are high-leverage outliers. In the location problem, M estimators such as the Huber and the bisquare bounded the influence of individual discrepant observations, but this is not the case in regression—if we admit the possibility of X-values with high leverage. High-leverage observations can force small residuals even when these observations depart from the pattern of the rest of the data.[14]

A key concept in assessing influence is the *breakdown point* of an estimator: The breakdown point is the fraction of arbitrarily "bad" data that the estimator can tolerate without being affected to an arbitrarily large extent. In the location problem, for example, the mean has a breakdown point of 0, because a *single* bad observation can change the mean by an arbitrary amount. The median, in contrast, has a breakdown point of 50%, because fully half the data can be bad without causing the median to become completely unstuck.[15] It is disquieting that in regression analysis, *all* M estimators have breakdown points of 0.

[14]For an illustration of this phenomenon, see Exercise 19.3.
[15]See Exercise 19.2.

There are regression estimators, however, that have breakdown points of nearly 50%. One such *bounded-influence* estimator is *least-trimmed-squares* (LTS) regression.[16]

Return to the fitted regression model $Y_i = \mathbf{x}_i'\mathbf{b} + E_i$, ordering the squared residuals from smallest to largest:[17] $(E^2)_{(1)}, (E^2)_{(2)}, \ldots, (E^2)_{(n)}$. Then, select $\mathbf{b}$ to minimize the sum of the smaller "half" of the squared residuals—that is,

$$\sum_{i=1}^{m} (E^2)_{(i)} \tag{19.4}$$

for $m = \lfloor (n + k + 2)/2 \rfloor$ (where the "floor" brackets indicate rounding down to the next smallest integer).

The LTS criterion is easily stated, but the LTS estimate is not so easily computed. One approach is to consider all subsets of observations of size $k + 1$ for which the vectors $\mathbf{x}_i'$ are distinct. Let the $(k + 1) \times (k + 1)$ model matrix for a particular such subset be represented as $\mathbf{X}^*$. Because the rows of $\mathbf{X}^*$ are all different, it is almost surely the case that the matrix $\mathbf{X}^*$ is of full rank, and we can compute the regression coefficients for this subset as $\mathbf{b}^* = \mathbf{X}^{*-1}\mathbf{y}^*$ (where $\mathbf{y}^*$ contains the corresponding entries of the response vector).[18] For each such subset, we compute the LTS criterion in Equation 19.4 and take as the LTS estimator $\mathbf{b}_{\text{LTS}}$ the value of $\mathbf{b}^*$ that minimizes this criterion.

If there are no repeated rows in the model matrix $\mathbf{X}$, then the number of subsets of observations of size $k + 1$ is

$$\binom{n}{k+1} = \frac{n!}{(n-k-1)!(k+1)!}$$

which is a very large number unless n is small. Even with highly efficient computational methods, it quickly becomes impractical, therefore, to find the LTS estimator by this approach. But we can compute a close approximation to $\mathbf{b}_{\text{LTS}}$ by randomly sampling many (but not unmanageably many) subsets of observations and minimizing the LTS criterion over the sampled subsets.

In the case of Duncan's occupational prestige regression, it is feasible to compute *all* subsets of size $k + 1 = 3$ of the $n = 45$ observations, of which there are

$$\binom{45}{3} = \frac{45!}{42!3!} = 14{,}190$$

The LTS estimates, it turns out, are similar to the bisquare estimates given in the previous section (Table 19.2 on page 594):

$$\widehat{\text{Prestige}} = -5.764 + 0.8023 \times \text{Income} + 0.4098 \times \text{Education}$$

[16]The LTS estimator, the *MM* estimator introduced below, and other bounded-influence estimators in regression are described in detail by Rousseeuw and Leroy (1987).

[17]Lest the notation appear confusing, note that it is the *squared* residuals E_i^2 that are ordered from smallest to largest, *not* the residuals E_i themselves.

[18]See Exercise 19.4.

Unlike the M estimator of location, the M estimator in regression is vulnerable to high-leverage observations. Bounded-influence estimators limit the impact of high-leverage observations. One such bounded-influence estimator is LTS, which selects the regression coefficients to minimize the smaller "half" of the squared residuals, $\sum_{i=1}^{m}(E^2)_{(i)}$ (where $m = \lfloor (n+k+2)/2 \rfloor$). The LTS estimator can be computed by calculating the regression coefficients for all subsets of observations of size $k+1$ and selecting the regression coefficients from the subset that minimizes the LTS criterion. If there are too many such subsets, then a manageable number can be sampled randomly.

LTS and other bounded-influence estimators are not a panacea for linear-model estimation, because they can give unreasonable results for some data configurations.[19] As well, the LTS estimator has much lower efficiency than the M estimators that we considered if the errors are in fact normal.

The latter problem can be addressed by combining bounded-influence estimation with M estimation, producing a so-called *MM* estimator, which retains the high breakdown point of the bounded-influence estimator and the high efficiency under normality of the M estimator. The *MM* estimator uses a bounded-influence estimator for start values in the computation of an M estimate and also to estimate the scale of the errors. For example, starting with the LTS estimator of the Duncan regression and following with the bisquare estimator yields the *MM* estimates

$$\widehat{\text{Prestige}} = -7.490 + 0.8391 \times \text{Income} + 0.3935 \times \text{Education}$$

The *MM* estimator combines the high breakdown point of bounded-influence regression with the high efficiency of M estimation for normally distributed errors. The *MM* estimator uses start values and a scale estimate obtained from a preliminary bounded-influence regression.

19.3 Quantile Regression

Quantile regression, due to Koenker and Bassett (1978), is a conceptually straightforward generalization of LAV regression. As I have explained, LAV regression estimates the conditional median (i.e., 50th percentile) of the response variable as a function of the explanatory variables. Quantile regression extends this approach to estimating other conditional quantiles of the response, such as the quartiles.

The LAV criterion in linear regression is written most directly as

$$\sum_{i=1}^{n} \rho_{\text{LAV}}(Y_i - \mathbf{x}_i'\mathbf{b}) \equiv \sum_{i=1}^{n} |Y_i - \mathbf{x}_i'\mathbf{b}|$$

The LAV estimator, $\mathbf{b}_{\text{LAV}}$, is the value of $\mathbf{b}$ that minimizes this criterion. An equivalent expression, the motivation for which will become clear presently, is

[19] See Stefanski (1991).

$$\sum_{i=1}^{n} \rho_{\text{LAV}}(Y_i - \mathbf{x}_i'\mathbf{b}) = 0.5 \times \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) < 0} |Y_i - \mathbf{x}_i'\mathbf{b}| + 0.5 \times \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) > 0} |Y_i - \mathbf{x}_i'\mathbf{b}|$$

that is, the LAV criterion consists of two components: The first component includes observations producing negative residuals and the second, observations producing positive residuals; residuals in these two classes are weighted *equally*.

Koenker and Bassett show that estimating the conditional q quantile (where $0 < q < 1$) is equivalent to minimizing

$$\sum_{i=1}^{n} \rho_q(Y_i - \mathbf{x}_i'\mathbf{b}) = q \times \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) < 0} |Y_i - \mathbf{x}_i'\mathbf{b}| + (1 - q) \times \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) > 0} |Y_i - \mathbf{x}_i'\mathbf{b}|$$

(i.e., a sum of *differentially weighted* negative and positive residuals) and that, furthermore, finding the value $\mathbf{b} = \mathbf{b}_q$ that minimizes this criterion is a straightforward linear programming problem.[20] They proceed to derive the asymptotic covariance matrix of the estimated quantile regression coefficients as[21]

$$V(\mathbf{b}_q) = \sigma_q^2 (\mathbf{X}'\mathbf{X})^{-1}$$

where

$$\sigma_q^2 \equiv \frac{q(1-q)}{p[P^{-1}(q)]}$$

Here, $p(\cdot)$ is the probability density function for the error distribution, and $P^{-1}(\cdot)$ is the quantile function for the errors (supposing, as may not be the case, that the errors are identically distributed). Thus, $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.[22] Note that σ_q^2 plays the same role as the error variance σ_ε^2 does in the formula for the covariance matrix of the least-squares estimates.[23] In applications, σ_q^2 is estimated from the distribution of the residuals.

Quantile regression estimates a linear model for the conditional quantile q of the response variable by minimizing the criterion

$$\sum_{i=1}^{n} \rho_q(Y_i - \mathbf{x}_i'\mathbf{b}) = \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) < 0} q \times |Y_i - \mathbf{x}_i'\mathbf{b}| + \sum_{i:(Y_i - \mathbf{x}_i'\mathbf{b}) > 0} (1 - q) \times |Y_i - \mathbf{x}_i'\mathbf{b}|$$

The asymptotic covariance matrix of the quantile regression estimator $\mathbf{b}_q$ is $V(\mathbf{b}_q) = \sigma_q^2 (\mathbf{X}'\mathbf{X})^{-1}$, where $\sigma_q^2 \equiv q(1-q)/\{p[P^{-1}(q)]\}$ and $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.

[20] Linear programming is a common type of optimization problem, for which there are well-understood and efficient methods. See, for example, Gass (2003).

[21] Koenker and Bassett (1978) also give exact finite-sample results, but these are too computationally demanding to prove useful in practice. An alternative to using the asymptotic standard errors is to base inference for quantile regression on the bootstrap, as described in Chapter 21.

[22] See the formula for the standard error of an order statistic given in Equation 3.4 (page 39).

[23] See Section 9.3.1.

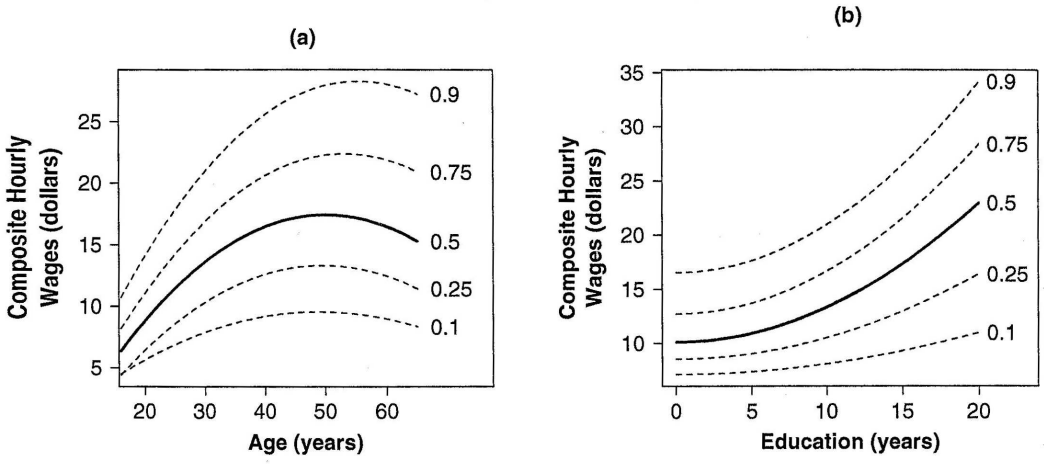

Figure 19.6 Effect displays for (a) age and (b) education in the quantile regression of wages on these variables and sex in the SLID data. In each case, the conditional .1, .25, .5, .75, and .9 quantiles are estimated. To construct each effect display, the other quantitative explanatory variable is set to its median value in the data, and the dummy regressor for sex is set to 0.5.

Figure 19.6 illustrates quantile regression by applying it to data from the Canadian Survey of Labour and Income Dynamics (SLID). I previously fit a model in which the log of the composite hourly wage rate for individuals in the sample was regressed on a dummy variable for sex, a quadratic in age, and the square of education.[24] For example, the estimated regression equation for the conditional median (i.e., the .5 quantile) is

$$\begin{aligned}\widehat{\text{Median Wages}} = &\underset{(0.69)}{-13.44} + \underset{(0.216)}{3.066\times\text{Male}} + \underset{(0.0485)}{0.9564\times\text{Age}} \\ &\underset{(0.000679)}{- 0.009567\times\text{Age}^2} + \underset{(0.00157)}{.03213\times\text{Education}^2}\end{aligned}$$

Asymptotic standard errors are in parentheses below the coefficients. Note in Figure 19.6 how the regression quantiles spread apart at higher values of age and education, where the median level of wages is relatively high, and how, for the most part, the conditional distribution of wages is positively skewed, with the upper quantiles more spread out than the lower ones. These characteristics, recall, motivated the log transformation of wages in least-squares regressions for these data.

Quantile regression is an attractive method not only because of its robustness relative to least squares but also because of its simple interpretation and because of its focus on the *whole*

[24]See Section 12.3. There is, however, a subtle change here: We were careful on log-transforming wages to make sure that the form in which age and education entered the model adequately captured the dependence of the conditional mean response on these explanatory variables. Because the log transformation is not linear, a quadratic in age and the square of education may not be appropriate for the conditional median of *untransformed* wages. I could, of course, compute instead the quantile regression for *log* wages (and I invite the reader to do so), but I wanted to illustrate how quantile regression can reveal asymmetry and nonconstant spread in the conditional distribution of the response.

conditional distribution of the response. Moreover, quantile regression extends naturally beyond linear models, for example, to nonlinear regression and to nonparametric regression.[25]

19.4 Robust Estimation of Generalized Linear Models

The maximum-likelihood or quasi-likelihood estimating equations for a generalized linear model can be written as

$$\sum_{i=1}^{n} \frac{1}{a_i}(Y_i - \mu_i) \underset{(k+1\times 1)}{\mathbf{x}_i} = \underset{(k+1\times 1)}{\mathbf{0}} \tag{19.5}$$

where Y_i is the response variable for the ith of n observations, $\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$ is the conditional expectation of the response given the values of the regressors $\mathbf{x}_i'$ for observation i, $\boldsymbol{\beta}$ is the parameter vector of $k+1$ regression coefficients to be estimated, and $g^{-1}(\cdot)$ is the inverse of the link function (i.e., the mean function) for the model. The constants a_i depend on the distributional family used to model the response; for example, for the Gaussian family $a_i = 1$, while for the binomial family $a_i = 1/n_i$ (the inverse of the number of binomial trials).[26] Because it depends directly on the difference between the observed and fitted response, the maximum-likelihood or quasi-likelihood estimator based on these estimating equations is generally not robust.

Cantoni and Ronchetti (2001) suggest replacing Equation 19.5 by estimating equations of the form

$$\sum_{i=1}^{n} \psi(Y_i, \mu_i) = \mathbf{0} \tag{19.6}$$

where $\psi(\cdot)$ is selected to produce high resistance to outliers. Equation 19.6 is a generalization of the estimating equations for M estimators in linear regression (Equation 19.3 on page 593). Bounded influence is achieved by down-weighting high-leverage observations: The weights employed are the product of (1) weights measuring the discrepancy between the observed and fitted response and (2) weights accounting for the leverage of the observations.[27] The details are beyond the scope of this presentation and are developed in Cantoni and Ronchetti's (2001) study.

Because the response variable in a binomial generalized linear model is bounded by 0 and 1, it is rare (but not impossible) to find highly influential observations in a logit or probit regression. GLMs for count data and for non-normal continuous responses are another matter, and robust estimation for these models is potentially useful. My limited experience with Cantoni

[25] See Koenker (2005).

[26] See Section 15.3.

[27] A simple choice of leverage-based weights is $\sqrt{1-h_i}$, where h_i is the hat-value for the ith observation (see Section 11.2). A higher breakdown point can be achieved, however, by using a robust covariance matrix for the Xs to judge the unusualness of observations in the X-space; using a robust covariance matrix is not sensible when the model matrix includes dummy regressors or other contrasts. Applied to *linear* regression, bounded-influence estimators using weights based on the product of leverage and discrepancy are called *GM* (*generalized-M*) *estimators* (see Rousseeuw & Leroy, 1987, Chapter 1).

and Ronchetti's estimator, however, suggests that it is not entirely reliable for detecting and discounting influential data.[28]

> Robust bounded-influence estimators for generalized linear models can be obtained by replacing the usual maximum-likelihood or quasi-likelihood estimating equations for GLMs by $\sum_{i=1}^{n} \psi(Y_i, \mu_i) = \mathbf{0}$, where $\psi(\cdot)$ is selected to produce high resistance to outliers. Bounded influence is achieved by down-weighting observations that have large residuals or large leverage.

19.5 Concluding Remarks

A final caution concerning robust regression: Robust estimation is not a substitute for close examination of the data. Although robust estimators can cope with heavy-tailed error distributions and outliers, they cannot correct nonlinearity, for example. Indeed, one use of robust estimation is to employ it as a routine diagnostic for unusual data in small- to medium-size samples, comparing the results obtained for a robust estimator with those of least-squares regression and investigating when the two estimators produce substantially different estimates (see, e.g., the discussion of the final weights for the Duncan regression displayed in Figure 19.5 on page 595).

As I have pointed out with respect to quantile regression, robust estimators can be extended to other settings. For example, it is a simple matter, and indeed common, to employ M estimator "robustness weights" in local-polynomial nonparametric regression, multiplying these weights by the neighborhood weights for the usual local-polynomial estimator, thereby rendering the local-polynomial estimator resistant to outliers.[29]

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 19.1. *Prove that the median minimizes the least-absolute-values objective function:

$$\sum_{i=1}^{n} \rho_{\text{LAV}}(E_i) = \sum_{i=1}^{n} |Y_i - \widehat{\mu}|$$

Exercise 19.2. Breakdown: Consider the contrived data set

$$Y_1 = -0.068 \quad Y_2 = -1.282 \quad Y_3 = 0.013 \quad Y_4 = 0.141$$
$$Y_5 = -0.980$$

[28] See, for example, Exercise 19.5.
[29] See Chapter 18.

(an adaptation of the data used to construct Figure 19.1). Show that more than two values must be changed to influence the median of the five values to an arbitrary degree. (Try, e.g., to make the first two values progressively and simultaneously larger, graphing the median of the altered data set against the common value of Y_1 and Y_2; then, do the same for the first three observations.)

Exercise 19.3. The following contrived data set (discussed in Chapter 3) is from Anscombe (1973):

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

(a) Graph the data and confirm that the third observation is an outlier. Find the least-squares regression of Y on X, and plot the least-squares line on the graph.
(b) Fit a robust regression to the data using the bisquare or Huber M estimator. Plot the fitted regression line on the graph. Is the robust regression affected by the outlier?
(c) Omitting the third observation $\{13, 12.74\}$, the line through the rest of the data has the equation $Y = 4 + 0.345X$, and the residual of the third observation from this line is 4.24. (Verify these facts.) Generate equally discrepant observations at X-values of 23 and 33 by substituting these values successively into the equation $Y = 4 + 0.345X + 4.24$. Call the resulting Y values Y_3' and Y_3''. Redo parts (a) and (b), replacing the third observation with the point $\{23, Y_3'\}$. Then, replace the third observation with the point $\{33, Y_3''\}$. What happens?
(d) Repeat part (c) using the LTS bounded-influence estimator. Do it again with the *MM* estimator.

Exercise 19.4. Computing the LTS estimator: Why is it almost surely the case that the $(k+1)\times(k+1)$ matrix $\mathbf{X}^*$, with rows selected from among those of the complete model matrix $\mathbf{X}$, is of full rank when all its rows are different? (Put another way, how is it possible that $\mathbf{X}^*$ would *not* be of full rank?) Thinking in terms of the $(k+1)$-dimensional scatterplot of Y against $X_1, \ldots, X_k$, what does the hyperplane defined by $\mathbf{b}^* = \mathbf{X}^{*-1}\mathbf{y}^*$ represent?

Exercise 19.5. In Chapter 15, I fit a Poisson regression of number of interlocks on assets, nation of control, and sector for Ornstein's Canadian interlocking-directorate data. The results from this regression are given in Table 15.3 (page 428). Influential-data diagnostics (see, e.g.,

Figure 15.7 on page 456) suggest that the first observation in the data set is quite influential; in particular, the coefficient of assets changes considerably when the first observation is removed. Perform a robust Poisson regression for this model. How do the results compare to removing the first observation from the data set? (Recall, however, that the influence of the first observation depends on unmodeled nonlinearity in the relationship between interlocks and assets—a problem that I ultimately addressed in Chapter 15 by log-transforming assets.)

Summary

- Robust M estimators of location, for the parameter μ in the simple model $Y_i = \mu + \varepsilon_i$, minimize the objective function

$$\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \widehat{\mu})$$

  selecting $\rho(\cdot)$ so that the estimator is relatively unaffected by outlying values. Two common choices of objective function are the Huber and the biweight (or bisquare).
- The sensitivity of an M estimator to individual observations is expressed by the influence function of the estimator, which has the same shape as the derivative of the objective function, $\psi(E) \equiv \rho'(E)$.
- An estimating equation for $\widehat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\widehat{\mu}$) to 0, obtaining $\sum_{i=1}^{n} \psi(Y_i - \widehat{\mu}) = 0$. The simplest procedure for solving this estimating equation is by iteratively reweighted means. Defining the weight function as $w(E) \equiv \psi(E)/E$, the estimating equation becomes $\sum_{i=1}^{n}(Y_i - \widehat{\mu})w_i = 0$, from which $\widehat{\mu} = \sum w_i Y_i / \sum w_i$. Starting with an initial estimate $\widehat{\mu}^{(0)}$, initial weights are calculated, and the value of $\widehat{\mu}$ is updated. This procedure continues iteratively until the value of $\widehat{\mu}$ converges.
- M estimation for the regression model $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ is a direct extension of M estimation of location: We seek to minimize an objective function of the regression residuals:

$$\sum_{i=1}^{n} \rho(E_i) = \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i'\mathbf{b})$$

  Differentiating the objective function and setting the derivatives to 0 produces the estimating equations

$$\sum_{i=1}^{n} \psi(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i' = \mathbf{0}$$

- Using the weight function, the estimating equations can be written as

$$\sum_{i=1}^{n} w_i(Y_i - \mathbf{x}_i'\mathbf{b})\mathbf{x}_i' = \mathbf{0}$$

  The solution of the estimating equations then follows by weighted least squares:

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

where $\mathbf{W}$ is the diagonal matrix of weights. The method of iterated weighted least squares starts with initial estimates $\mathbf{b}^{(0)}$, calculates initial residuals from these estimates, and calculates initial weights from the residuals. The weights are used to update the parameter estimates, and the procedure is iterated until it converges.

- Unlike the M estimator of location, the M estimator in regression is vulnerable to high-leverage observations. Bounded-influence estimators limit the effect of high-leverage observations. One such bounded-influence estimator is LTS, which selects the regression coefficients to minimize the smaller "half"of the squared residuals $\sum_{i=1}^{m}(E^2)_{(i)}$ (where $m = \lfloor (n+k+2)/2 \rfloor$). The LTS estimator can be computed by calculating the regression coefficients for all subsets of observations of size $k+1$ and selecting the regression coefficients from the subset that minimizes the LTS criterion. If there are too many such subsets, then a manageable number can be sampled randomly.
- The *MM* estimator combines the high breakdown point of bounded-influence regression with the high efficiency of M estimation for normally distributed errors. The *MM* estimator uses start values and a scale estimate obtained from a preliminary bounded-influence regression.
- Quantile regression estimates a linear model for the conditional quantile q of the response variable by minimizing the criterion

$$\sum_{i=1}^{n}\rho_q(Y_i - \mathbf{x}_i'\mathbf{b}) = \sum_{i:\,(Y_i-\mathbf{x}_i'\mathbf{b})<0} q\times\left|Y_i - \mathbf{x}_i'\mathbf{b}\right| + \sum_{i:\,(Y_i-\mathbf{x}_i'\mathbf{b})>0}(1-q)\times\left|Y_i - \mathbf{x}_i'\mathbf{b}\right|$$

  The asymptotic covariance matrix of the quantile regression estimator $\mathbf{b}_q$ is $V(\mathbf{b}_q) = \sigma_q^2(\mathbf{X}'\mathbf{X})^{-1}$, where $\sigma_q^2 \equiv q(1-q)/\{p[P^{-1}(q)]\}$, and $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.
- Robust bounded-influence estimators for generalized linear models can be obtained by replacing the usual maximum-likelihood or quasi-likelihood estimating equations for GLMs by $\sum_{i=1}^{n}\psi(Y_i, \mu_i) = \mathbf{0}$, where $\psi(\cdot)$ is selected to produce high resistance to outliers. Bounded influence is achieved by down-weighting observations that have large residuals or large leverage.

Recommended Reading

- In a volume on robust and exploratory methods, edited by Hoaglin, Mosteller, and Tukey (1983), Goodall (1983) presents a high-quality, readable treatment of M estimators of location.
- A fine chapter by Li on M estimators for regression appears in a companion volume (Hoaglin, Mosteller, & Tukey, 1985).
- Another good source on M estimators is Wu (1985).
- Rousseeuw and Leroy's (1987) book on robust regression and outlier detection emphasizes bounded-influence, high-breakdown estimators.
- Andersen (2007) presents a broad and largely accessible overview of methods of robust regression, including a discussion of robust estimation for generalized linear models.
- Koenker (2005) offers an extensive treatment of quantile regression by the originator of the method.