CHAPTER 25

Robust Regression

Why Use Robust Regression?

When we perform least squares regression using n observations, for a p-parameter model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we make certain idealized assumptions about the vector of errors $\boldsymbol{\epsilon}$, namely, that it is distributed $N(\mathbf{0}, \mathbf{I}\sigma^2)$. In practice, departures from these assumptions occur. If the departures are serious, we hope to spot them in the behavior of the residuals and so be led to make suitable adjustments to the model and/or the variables' metrics. For example, we may transform the response variable or one or more of the predictors, or adjust the model by adding higher-order terms. For many sets of data, the departures, if they exist at all, are not serious enough for corrective actions, and we proceed with the analysis in the usual way.

If our analysis seems to point to the errors having a non-normal distribution, we might consider a robust regression method, particularly in cases where the error distribution is heavier-tailed then the normal, that is, has more probability in the tails than the normal. Such heavier-tailed distributions are likely to generate more "large" errors than the normal. A least squares analysis weights each observation equally in getting parameter estimates. Robust methods enable the observations to be weighted unequally. Essentially, observations that produce large residuals are down-weighted by a robust estimation method. A number of methods are available. In general, robust regression methods require much more computing than least squares, and also require some assumptions to be made about the down-weighting procedure to be employed.

25.1. LEAST ABSOLUTE DEVIATIONS REGRESSION (L_1 REGRESSION)

When we can make a specific (non-normal) assumption about the errors, we can apply maximum likelihood methods directly. As already discussed in Section 5.1, the assumption of a double exponential error distribution

$$(2\sigma)^{-1} \exp\{-|\epsilon|/\sigma\} \qquad (-\infty \leq \epsilon \leq \infty)$$

leads to minimizing

$$\sum_{i=1}^{n} |\epsilon_i|,$$

the sum of absolute errors. This is also called L_1 regression (or L_1-norm regression), the subscript 1 referring to the power used in the function to be minimized. This

method gives less weight to larger errors than the least squares method, where the error distribution is

$$(2\pi\sigma^2)^{-1/2} \exp\{-\epsilon^2/(2\sigma^2)\}$$

and the function minimized is the sum of squares of errors, namely,

$$\sum_{i=1}^{n} \epsilon_i^2.$$

Thus L_1 regression is more robust against large errors than least squares; least squares is sometimes called L_2 regression (or L_2-norm regression). There exist also L_p regression methods that minimize

$$\sum_{i=1}^{n} |\epsilon_i|^p.$$

It has been suggested (by Forsythe, 1972, on the basis of a 400-replicates Monte Carlo study) that a value of $p = 1.5$ could be a good general choice. It led to estimates that were no worse than 95% as efficient as least squares when the errors were actually normal. Efficiency was defined as the ratio (mean square error for least squares)/(mean square error for power p), and Forsythe's Monte Carlo experiments used $p = 1.25$, 1.50, and 1.75 on sets of straight line regression data with errors generated from a normal distribution contaminated with another off-center normal to produce a skewed distribution of errors. See Forsythe (1972, p. 160) for details.

25.2. *M*-ESTIMATORS

M-Estimators are "maximum likelihood type" estimators. Suppose the errors are independently distributed and all follow the same distribution, $f(\epsilon)$. Then the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}$, which maximizes the quantity

$$\prod_{i=1}^{n} f(Y_i - \mathbf{x}_i'\boldsymbol{\beta}), \tag{25.2.1}$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$, $i = 1, 2, \ldots, n$, in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Equivalently, the MLE of $\boldsymbol{\beta}$ maximizes

$$\sum_{i=1}^{n} \ln f(Y_i - \mathbf{x}_i'\boldsymbol{\beta}). \tag{25.2.2}$$

As we have seen, this leads to minimizing the sum of squares function

$$\sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \tag{25.2.3}$$

in the normal case. In the double exponential case we minimize

$$\sum_{i=1}^{n} |Y_i - \mathbf{x}_i'\boldsymbol{\beta}|. \tag{25.2.4}$$

We can extend this idea as follows. Suppose $\rho(u)$ is a defined function of u (examples soon) and suppose s is an estimate of scale (not necessarily the usual least squares estimate). We define a robust estimator as one that minimizes

$$\sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \sum_{i=1}^{n} \rho\left\{\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right\}. \tag{25.2.5}$$

We see that if $\rho(u) = u^2$, the criterion minimized is the same as that of (25.2.3), while if $\rho(u) = |u|$ we get (25.2.4). So in these specific cases, the form of ρ and the underlying distribution are specifically related. In fact, the *knowledge* of an appropriate distribution for the errors could tell us what $\rho(u)$ to use.

Here, we encounter the first practical difficulty in using robust regression. In general, if we do not know what distribution to assume for the errors, we are not led naturally to a $\rho(u)$. What we have essentially in the literature are a number of suggestions for $\rho(u)$ that have worked out well in specific studies. It is not clear which should be used for a given set of data. All we can do is look at the weighting characteristics produced by $\rho(u)$ and choose accordingly, all else being equal.

Table 25.1 (*a*)–(*f*) shows some of the suggestions given in the literature for $\rho(u)$. Choices of specific values for the constants a, b, and c need to be made; values shown are suggested in the literature. [Row (*g*), attached here for convenience, shows an appropriate weight function $w(u)$, for the situation where it is assumed that the underlying error distribution is a t-distribution with f degrees of freedom; see Box and Draper, 1987, pp. 79–90.]

In Eq. (25.2.5), the role of u is played by the scaled residual $(Y_i - \mathbf{x}_i'\boldsymbol{\beta})/s$. We shall soon see, from Figure 25.1, that the least squares method gives the highest weight of 1 in (25.2.5) to larger scaled residuals, and the purpose of the various $\rho(u)$ functions is to (comparatively) down-weight larger scaled residuals in various ways, compared with least squares. We discuss this more in a moment. First we examine how the estimation procedure is carried out.

The *M*-Estimation Procedure

We need to minimize Eq. (25.2.5) with respect to the parameters β_j, $j = 0, 1, 2, \ldots, k$, say. Differentiation of Eq. (25.2.5) with respect to each parameter leads to $p = k + 1$ equations of form

$$\sum_{i=1}^{n} x_{ij}\psi\left\{\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right\} = 0, \qquad j = 0, 1, 2, \ldots, k, \tag{25.2.6}$$

where $\psi(u)$ is the partial derivative $\partial\rho/\partial u$ and x_{ij} is the jth entry of $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$. These equations do not have an explicit solution in general, and iterative numerical solution is necessary. We can follow Beaton and Tukey (1974, p. 151) and define weights

$$w_{i\beta} = \frac{\psi\{(Y_i - \mathbf{x}_i'\boldsymbol{\beta})/s\}}{(Y_i - \mathbf{x}_i'\boldsymbol{\beta})/s}, \qquad i = 1, 2, \ldots, n, \tag{25.2.7}$$

defining them to be 1 if it happens that $Y_i - \mathbf{x}_i'\boldsymbol{\beta} = 0$ exactly. Then (25.2.6) becomes

$$\sum_{i=1}^{n} x_{ij} w_{i\beta}(Y_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0, \qquad j = 0, 1, 2, \ldots, k, \tag{25.2.8}$$

TABLE 25.1. Some Functions that Have Been Suggested for *M*-Estimation

Criterion	$\rho(u)$	Corresponding Ranges of u	$\psi(u) = \frac{\partial \rho(u)}{\partial u}$	$w(u) = \psi(u)/u$	Suggested Parameter or Breakpoint Values (Tuning Constants) and Reference
(*a*) Least squares (normal errors)	$\frac{1}{2}u^2$	$-\infty \le u \le \infty$	u	1	Not applicable
(*b*) Huber's, with breakpoint $a > 0$	$\begin{cases} \frac{1}{2}u^2 \\ a\lvert u\rvert - \frac{1}{2}a^2 \end{cases}$	$-a \le u \le a$ $u \le -a$ and $u \ge a$	u a sign (u)	1 $a/\lvert u\rvert$	$a = 2$ (Huber, 1964)
(*c*) Ramsay's, with parameter $a > 0$	$\{1 - e^{-a\lvert u\rvert}(1 + a\lvert u\rvert)\}/a^2$	$-\infty \le u \le \infty$	$ue^{-a\lvert u\rvert}$ (maximum at a^{-1})	$e^{-a\lvert u\rvert}$	$a = 0.3$ (Ramsay, 1977)
(*d*) Andrew's wave, with breakpoint $a\pi > 0$	$\begin{cases} a\{1 - \cos(u/a)\} \\ 2a \end{cases}$	$-a\pi \le u \le a\pi$ $u \le -a\pi$ and $u \ge a\pi$	$\sin(u/a)$ 0	$\{\sin(u/a)\}/(u/a)$ 0	$a = 1.339$ (Andrews et al., 1972)
(*e*) Tukey's biweight, with breakpoint $a > 0$	$\begin{cases} \frac{1}{2}u^2 - \frac{u^4}{4a^2} \\ \frac{1}{4}a^2 \end{cases}$	$-a \le u \le a$ $u \le -a$ and $u \ge a$	$u(1 - u^2/a^2)$ 0	$1 - u^2/a^2$ 0	$5 \le a \le 6$ (Beaton and Tukey, 1974)
(*f*) Hampel's, with breakpoints a, b, $c > 0$	$\begin{cases} \frac{1}{2}u^2 \\ a\lvert u\rvert - \frac{1}{2}a^2 \\ a\dfrac{(c\lvert u\rvert - \frac{1}{2}u^2)}{c-b} - \dfrac{7a^2}{6} \\ a(b + c - a) \end{cases}$	$-a \le u \le a$ $-b \le u \le -a$ and $a \le u \le b$ $-c \le u \le -b$ and $b \le u \le c$ $u \le -c$ and $u \ge c$	u a sign (u) $\{c$ sign $(u) - u\}a/(c-b)$ 0	1 $a/\lvert u\rvert$ $\{c/\lvert u\rvert - 1\}a/(c-b)$ 0	$a = 1.7$, $b = 3.4$, $c = 8.5$ (Andrews et al., 1972, p. 14)
(*g*) Assume errors follow t_f distribution and apply maximum likelihood		$-\infty \le u \le \infty$		$\frac{f+1}{f+u^2}$	Not applicable. As $f \to \infty$, the t-distribution $\to$ normal distribution

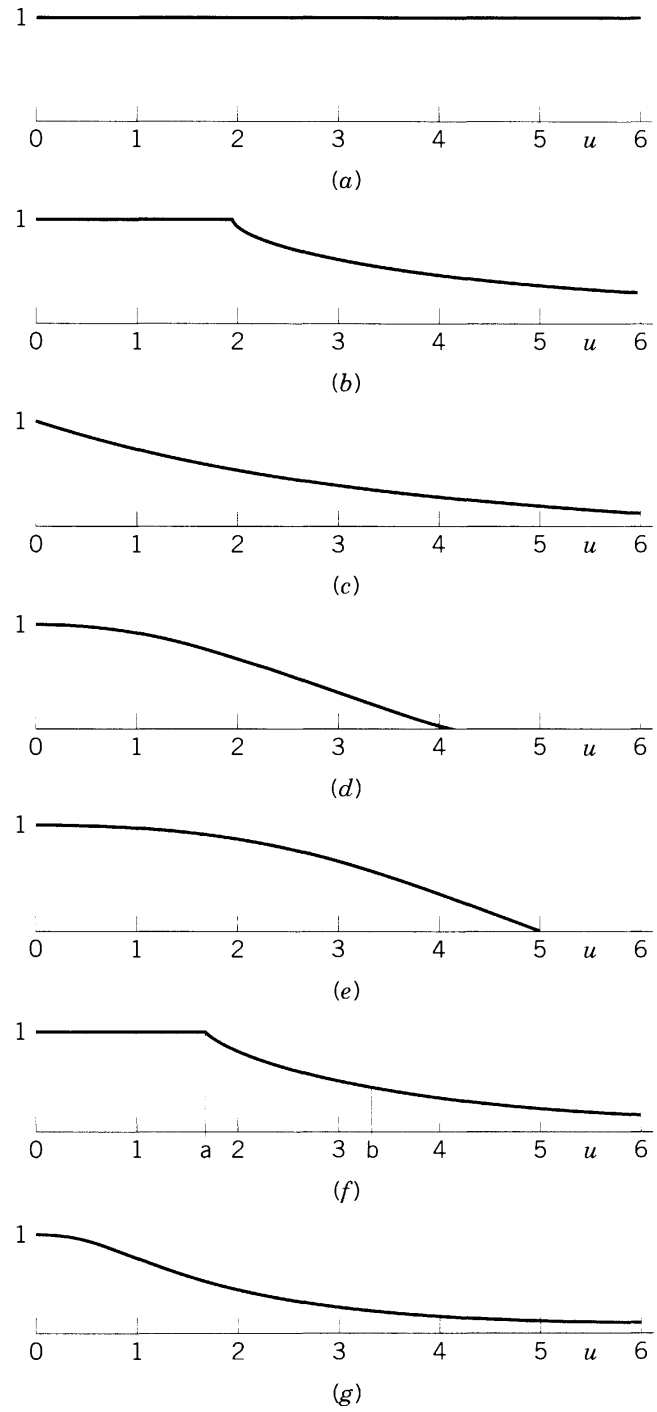

Figure 25.1. Plots of the weight functions $w(u) = u^{-1}\,\partial\rho(u)/\partial u$ for the $\rho(u)$ in Table 25.1. The plots arise from these choices of "tuning constants" from Table 25.1: (*b*) $a = 2$; (*c*) $a = 0.3$; (*d*) $a = 1.339$; (*e*) $a = 5$; (*f*) $a = 1.7$, $b = 3.4$, $c = 8.5$; (*g*) $f = 2$. Other choices provide different weight plots of essentially the same nature.

or

$$\sum_{i=1}^{n} x_{ij} w_{i\beta} \mathbf{x}_i' \boldsymbol{\beta} = \sum_{i=1}^{n} x_{ij} w_{i\beta} Y_i, \qquad j = 0, 1, 2, \ldots, k, \tag{25.2.9}$$

which we can write in matrix form as

$$\mathbf{X}'\mathbf{W}_\beta \mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}_\beta \mathbf{Y}, \tag{25.2.10}$$

where $\mathbf{W}_\beta$ = diagonal $(w_{1\beta}\ w_{2\beta}, \ldots, w_{n\beta})$. These equations are obviously of the form of generalized least squares normal equations. (More accurately, they are *weighted* least squares type equations because $\mathbf{W}_\beta$ is a diagonal matrix of weights.) The difficulty in solving them is that $\mathbf{W}_\beta$ depends on β. This is circumvented in a traditional way. We somehow get an initial estimate $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}$ (perhaps by using least squares), calculate the value $\mathbf{W}_0$, say, of $\mathbf{W}_\beta$ for that $\hat{\boldsymbol{\beta}}_0$, and solve (25.2.10) to get solution $\hat{\boldsymbol{\beta}}_1$. Using $\hat{\boldsymbol{\beta}}_1$ in $\mathbf{W}_\beta$ gets us $\mathbf{W}_1$. Then $\mathbf{W}_1$ is used to get $\hat{\boldsymbol{\beta}}_2$, and so on. Typically, convergence occurs quite quickly. (The procedure can also be halted after a selected number of steps, if desired.) Only a standard weighted least squares program is needed, as in MINITAB, for example. We can write the iterative solution as

$$\hat{\boldsymbol{\beta}}_{q+1} = (\mathbf{X}'\mathbf{W}_q\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_q\mathbf{Y}, \qquad q = 0, 1, \ldots, \tag{25.2.11}$$

and the procedure may be stopped when all the estimates change by less than some selected preset amount, say, 0.1% or 0.01%, or after a selected number of steps.

We see from (25.2.7) that the weights depend on the values of $[\partial\rho(u)/\partial u]/u$, where $u = (Y_i - \mathbf{x}_i'\boldsymbol{\beta})/s$. Thus we can look at the function $w(u) = [\partial\rho(u)/\partial u]/u$ for all the various choices of $\rho(u)$ in Table 25.1. Figure 25.1 shows plots of the weight functions $w(u) = [\partial\rho(u)/\partial u]/u$ for $u \geq 0$. (The $u \leq 0$ portion is symmetric with $u \geq 0$.) Obviously, choice of a particular $\rho(u)$ is basically a choice of what weight function the user wishes to assign to scaled residuals of various sizes. [One could even simply make up one's own weight function without defining a $\rho(u)$.]

The particular plots (b)–(g) shown in Figure 25.1 are based on specific assumptions that could be varied. For example, (f) assumes cutoff points $a = 1.7$, $b = 3.4$, and $c = 8.5$, at which point the weight becomes zero. If other a, b, c values were used, the diminishing-to-the-right three-part nature of (f) would remain, but the weight function would be somewhat compressed or expanded. Similarly, plot (b) is drawn with the change point at $u = 2$; this point could move left or right, and so on. The weight functions (d) and (e) are very similar in nature and could be adjusted to end at the same cutoff point. Moreover (e) is easier to deal with than (d), which involves trigonometric functions. Choice of the weight function to use, and of the constants (often called the *tuning constants*), is a second practical difficulty of robust regression. See, for example, Kelly (1996).

We next look at a couple of examples using robust regression. We recall that the weights are defined in terms of $u = (Y_i - \mathbf{x}_i'\boldsymbol{\beta})/s$ in (25.2.7), and require evaluation of a suitable scale factor s. One robust choice that has gained a measure of popularity is

$$s = \text{median}\ |e_i - \text{median}\ (e_i)|/0.6745, \tag{25.2.12}$$

which would provide a roughly unbiased estimator of $\text{sd}(Y_i) = \sigma$ if n were large and the errors were normally distributed. (This suggestion of Hampel is given, with a misprinted denominator, in Andrews et al., 1972, p. 12. It is also sometimes shown elsewhere with the denominator already inverted as 1.4826.) Several alternative scale factors have been proposed and discussed; see Hill and Holland (1977, pp. 830–831), for example. The actual choice for s *does* have an effect on the results; that is, the

robust regression parameter estimates are not invariant to choice of s. This is a third practical difficulty of robust estimation.

25.3. STEEL EMPLOYMENT EXAMPLE

Table 25.2 shows steel employment by country in thousands of people for the years 1974 and 1992. We propose to fit the model

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{25.3.1}$$

where Y = 1992 employment and X = 1974 employment. The least squares fit is

$$\hat{Y} = -0.314 + 0.4004X, \tag{25.3.2}$$

with $R^2 = 0.736$. Observations 1 and 4, with internally studentized residuals of 2.53 and −2.07, are designated "large." The corresponding externally studentized residuals are 5.34 and −2.83. We now make a robust fit using Huber's criterion with breakpoints at ±2.

The residuals are 39.4, 11.9, −19.9, −36.4, −2.3, −0.3, 6.3, −0.9, 1.7, 0.5 rounded to one decimal place. The median of the residuals is (0.5 − 0.3)/2 = 0.1 and the $|e_i - \text{median } (e_i)|$ values are thus 39.3, 11.8, 20.0, 36.5, 2.4, 0.4, 6.2, 1.0, 1.6, 0.4. These have median (6.2 + 2.4)/2 = 4.3. This gives $s = 4.3/0.6745 = 6.375$ if we use (25.2.12). Dividing this into the residuals gives u values of

$$6.2,\ 1.9,\ -3.1,\ -5.7,\ -0.4,\ 0.0,\ 1.0,\ -0.1,\ 0.3,\ 0.1. \tag{25.3.3}$$

Applying the weight formula of Table 25.1(b) with $a = 2$ results in initial weights of

$$0.323,\ 1,\ 0.645,\ 0.351,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1. \tag{25.3.4}$$

These weights are inserted in (25.2.10) and iterated upon, leading to the fit

$$\hat{Y} = 3.334 + 0.3205X \tag{25.3.5}$$

with final weights of 0.208, 0.711, and 0.462 for observations 1, 2, and 4 and unity for

TABLE 25.2. Steel Employment by Country in Europe, by Thousands, in 1974 and 1992

Country	1974	1992
Germany	232	132[a]
Italy	96	50
France	158	43
United Kingdom	194	41
Spain	89	33
Belgium	64	25
Netherlands	25	16
Luxembourg	23	8
Portugal	4	3
Denmark	2	1
Total	887	353

[a] Includes the former East Germany.

all others. The fitted least squares and robust fit lines are shown in Figure 25.2. The robust fit clearly down-weights heavily the outlying and influential first and fourth data points whose Cook's statistics are 2.54 and 0.85 in the original least squares fit. (The full listing is: 2.54, 0.02, 0.12, 0.85, 0, 0, 0.01, 0, 0, 0.)

Adjusting the First Observation

Prompted by the robust fit above, we now take note of the footnote in Table 25.2. Since the 1992 figure for Germany includes the former East Germany (whereas the 1974 figure does not), it may be too large. We can adjust it roughly by the factor $\frac{63}{80}$, approximately the ratio of West German and total German populations in 1992. This replaces 132 by $\frac{63}{80}$ (132) = 104. The new least squares fit is now

$$\hat{Y} = 2.803 + 0.3337X, \tag{25.3.6}$$

with $R^2 = 0.798$. Observations 1 and 4 have internally studentized residuals of 2.19 and -2.16. The externally studentized values are 3.22 and -3.11 and the ten Cook's statistics are, in observation order, 1.89, 0.07, 0.10, 0.92, 0, 0, 0.01, 0, 0, 0.01. Initial weights used for the Huber's criterion fit with breakpoints at ± 2 were calculated as in the previous fit, using (25.2.12) to get $s = 5.4855$, as

$$0.461, 0.721, 0.878, 0.414, 1, 1, 1, 1, 1, 1. \tag{25.3.7}$$

The robust fit that emerges is

$$\hat{Y} = 3.344 + 0.3205X \tag{25.3.8}$$

with final weights of 0.430, 0.711, and 0.462 for observations 1, 2, and 4 and unity for all others. Surprisingly, the robust fit is the same as before in (25.3.5), and the final weights of observations 2 and 4 are unchanged. The adjusted first observation is still heavily down-weighted but its weight has approximately doubled. The fitted least squares and robust fit lines are shown on Figure 25.3. On this scale they are not very distinguishable, in spite of the down-weighted observations. The adjustment of the largest Y appears to have been effective.

One might also question here whether the line should be forced to pass through the origin, since clearly if $X = 0$, $Y = 0$ necessarily. If this is done here, we get the no-intercept fitted model

$$\hat{Y} = 0.3516X, \tag{25.3.9}$$

the slope indicating the estimated reduction ratio of employment from 1974 to 1992

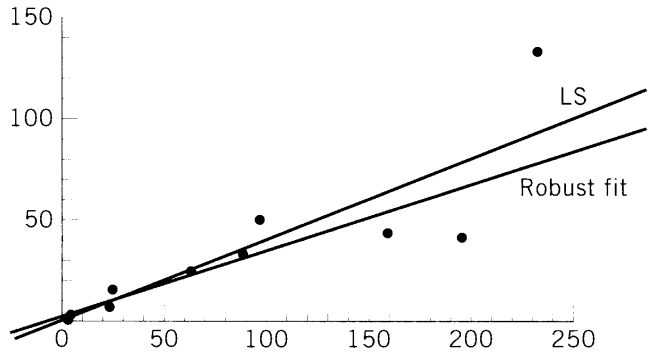

Figure 25.2. Least squares and robust fits to the original steel data.

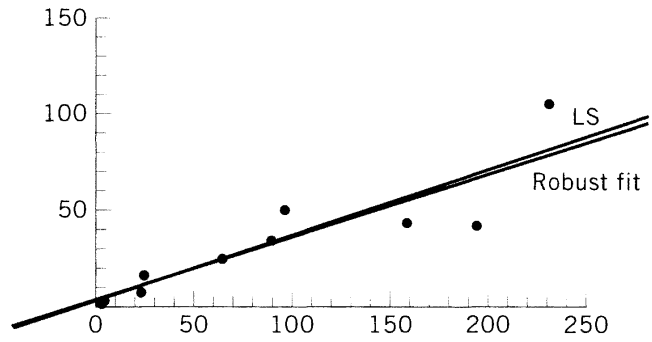

Figure 25.3. Least squares and robust fits to the adjusted steel data.

as about a third. This line is very close to the (two) lines in Figure 25.3 but passes through the origin and, at $X = 250$, lies merely 1.7 units higher than the least squares line and 4.4 units higher than the robust fit line.

Other Analyses Possible

There are several other robust analyses that could be made, using criteria other than Huber's from Table 25.1. Note that the weighting schemes used in both fits correspond well to the basic idea that the larger Y values might be subject to more error than the smaller Y values. This would not be surprising in data of this type. Our overall conclusion, however, is that any of the fits (25.3.6), (25.3.8), or (25.3.9) would be acceptable, all things considered. Perhaps (25.3.9) makes the most sense. The robust fits have greatly aided our view of how these data might be analyzed.

Standard Errors of Estimated Coefficients

Values for standard errors of the robust regression estimates can be obtained from the square roots of the diagonal entries in $(\mathbf{X}'\mathbf{W}_q\mathbf{X})^{-}s^2$, where q is set to its final value in the iterative procedure and s^2 is obtained from the final version of (25.2.12). A number of alternative calculations have also been proposed, however. For comments and additional references, see Hill and Holland (1977), Huber (1973), and Birch and Agard (1993).

25.4. TREES EXAMPLE

Table 25.3 shows tree data obtained by Tamra J. Burcar. The 25 observations (plotted in Figure 25.4) consist of values of

X = tree diameter in inches, measured at breast height (DBH),

Y = tree height in feet.

How well can we estimate height using a prediction equation based on tree diameter X?

We first carry out the least squares fit

$$\hat{Y} = 41.956 + 2.7836X, \tag{25.4.1}$$

with $R^2 = 0.633$. It is obvious that extrapolation of this equation below the data makes

TABLE 25.3. Tree Data, from Tamra J. Burcar, on Diameters (X) and Heights (Y) of 24 Trees

Observation Number	Y	X	Observation Number	Y	X
1	58	5.5	14	75	10.1
2	60	5.7	15	72	10.2
3	42	5.8	16	78	10.4
4	64	6.5	17	65	10.6
5	60	6.6	18	80	10.6
6	65	6.7	19	82	10.8
7	56	6.9	20	70	11.3
8	57	7.0	21	74	11.3
9	70	7.3	22	68	11.6
10	68	8.3	23	68	11.6
11	65	8.6	24	82	13.0
12	70	9.5	25	88	18.0
13	63	10.0			

no sense, as it predicts heights of above 40 feet on trees with close to zero diameter. It is also clear from the plot and the computer output that the third observation and the 25th are inconsistent with the rest of the data:

Observation Number:	3	25
Residual	−16.10	−4.06
Internally studentized residual	−2.76	−0.87
Externally studentized residual	−3.29	−0.86
Cook's statistic	0.44[a]	0.28[a]

[a] The other Cook's statistics range from 0 to 0.07.

Both observations 3 and 25 are influential, but only 3 has a large residual. Both these data points had already attracted suspicion. The small tree was clearly stunted and the large tree protruded above the tree line and may have been damaged by winds. These trees also stand out on the plot of Figure 25.4. All in all, there are ample

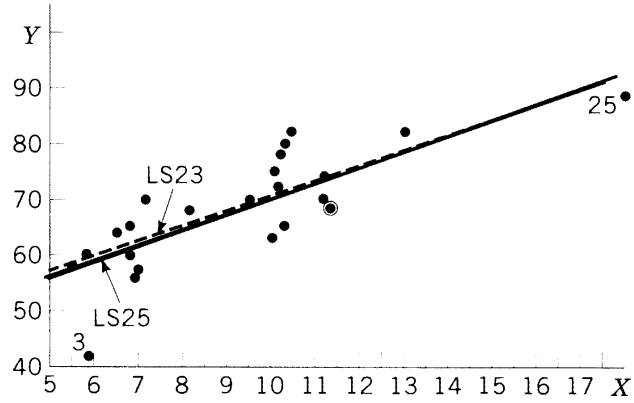

Figure 25.4. Tree data. Least squares fit to 25 points (LS25), to 23 points (LS 23), and a robust fit to 25 points (middle long-dashed line).

grounds for removing them entirely and fitting a straight line to the remaining data. However, we first procede with a robust analysis. We shall again use Huber's procedure with $a = 2$. Our estimate of scale, given via (25.2.12) using the least squares residuals, is

$$s = 4.1953/0.6745 = 6.22. \tag{25.4.2}$$

Only one observation, No. 3, is given a non-unit weight, of value 0.7726. Subsequently, iterations via (25.2.10) produce the robust fit

$$\hat{Y} = 42.872 + 2.7043X \tag{25.4.3}$$

with a weight 0.7377 for the third observation, and unity for all others. This line is compared with the least squares line (LS25) in Figure 25.4. It differs only slightly from it. The least squares line obtained by fitting to 23 data points, omitting the influential third and 25th observations, is

$$\hat{Y} = 44.353 + 2.6172X. \tag{25.4.4}$$

We see from Figure 25.4 that omission of the third point has allowed the intercept to rise very slightly, and that is the main change compared with the (quite similar) fits via least squares and robust estimation to all 25 observations. The robust fit tells us that observations 3 and 25 make only a slight difference compared to the least squares fit if all 25 observations are used. Thus the real decision is whether to use or discard observations 3 and 25. [Based on conversations with the scientist who provided these data, we would be inclined to discard them and use (25.4.4).] Again, making a robust fit to all 25 observations has helped our assessment of these data.

[If the least squares fit (25.4.4) is used as a "starter" for Huber's criterion with $a = 2$, the weights are all of size 1, so a robust fit to the 23 observations is not indicated.]

25.5. LEAST MEDIAN OF SQUARES (LMS) REGRESSION

Another robust regression method is the *least median of squares*, or LMS, method. This is given by minimizing, with respect to the elements of $\boldsymbol{\beta}$,

$$\underset{i}{\text{Median}}\,(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2.$$

For any given $\boldsymbol{\beta}$, we could find the median of the squared residuals; the solution is that $\boldsymbol{\beta}$ that produces the minimum such median. Geometrically, this method amounts to finding the narrowest strip (narrowest in the Y-axis direction) that covers "half" of the observations, where half means "the integer part of $n/2$, plus one." The LMS line lies at the center of this band. The computations are intricate and require a specially written program. See Rousseeuw and Leroy (1987), for example.

25.6. ROBUST REGRESSION WITH RANKED RESIDUALS (rreg)

There are a number of procedures of this type, which involve the minimization with respect to $\boldsymbol{\beta}$ of a weighted sum of the residuals. Suppose $a(\cdot)$ is a weight function (to be chosen) and $\mathbf{x}_i'$ is the ith row of the usual $\mathbf{X}$ matrix, $i = 1, 2, \ldots, n$. Then $Y_i -$

$\mathbf{x}_i'\boldsymbol{\beta}$ is the deviation of Y_i from the corresponding model function. Consider

$$D(\boldsymbol{\beta}) = \sum_{i=1}^{n} a\{R(Y_i - \mathbf{x}_i'\boldsymbol{\beta})\}\{Y_i - \mathbf{x}_i'\boldsymbol{\beta}\}, \tag{25.6.1}$$

where $R(Y_i - \mathbf{x}_i'\boldsymbol{\beta})$ is the rank of the ith deviation. Minimization with respect to $\boldsymbol{\beta}$ of this dispersion function will give a robust rank estimator of $\boldsymbol{\beta}$. There are many possible choices for the function $a(\cdot)$. One possible choice is the Wilcoxon score function

$$a(i) = 12^{1/2}\left\{\frac{i}{n+1} - \frac{1}{2}\right\} \tag{25.6.2}$$

for which

$$D_W(\boldsymbol{\beta}) = 12^{1/2}\sum_{i=1}^{n}\left\{\frac{i}{n+1} - \frac{1}{2}\right\}\{Y_i - \mathbf{x}_i'\boldsymbol{\beta}\}, \tag{25.6.3}$$

where the notation i in (25.6.2) and (25.6.3) *now* implies that the residuals are in rank order. We first illustrate with a small constructed example. [The factor $12^{1/2}$ in (25.6.3) can be dropped in the fitting process.]

Example 1. To the five data points $(X, Y) = (2, 2), (3, 3), (5, 4), (9, 6), (12, 7)$, we fit a straight line. Minimizing (25.6.3), using the rreg option in MINITAB, for example, gives

$$\hat{Y}_r = 1.25 + 0.5X.$$

The least squares fit (which uses weights of 1 for all the observations) is

$$\hat{Y} = 1.3701 + 0.48870X.$$

The best horizontal straight line fit is $\hat{Y} = \overline{Y} = 4.4$. (This is not a good fit but provides another fit to compare.) The weights $(i/6 - \frac{1}{2})$ are $(-\frac{1}{3}, -\frac{1}{6}, 0, \frac{1}{3}, \frac{1}{6})$. Table 25.4 shows the three sets of residuals, how the weights would be allocated to these residuals, if rreg were applied, and the values of the statistic (25.6.3) for the three cases. We can see from this example that, since the more extreme residuals receive larger (Wilcoxon) weights in computing $D_W(\boldsymbol{\beta})$ and that the $\boldsymbol{\beta}$ is chosen to minimize $D_W(\boldsymbol{\beta})$, this procedure will tend to reduce some larger residuals to attain the fit. Since the least squares fit has the smallest residual sum of squares, other lower weighted residuals will necessarily increase in modulus in the rreg fit compared to the least squares fit.

TABLE 25.4. Residuals and Wilcoxon Weights for Three Fits, Example 1

		rreg		Least Squares		$\hat{Y} = \overline{Y}$	
X	Y	Residuals	Weights	Residuals	Weights	Residuals	Weights
2	2	−0.25	−0.33	−0.35	−0.16	−2.4	−0.33
3	3	0.25	0.16	0.16	0.	−1.4	−0.16
5	4	0.25	0.	0.19	0.16	−0.4	0.
9	6	0.25	0.33	0.23	0.33	1.6	0.16
12	7	−0.25	−0.16	−0.23	−0.33	2.6	0.33
Cross-product/$12^{1/2}$		0.25		0.25942		2.167	
Residual SS		0.3125	—	0.29	—	17.2	—

Example 2. Steam data (Appendix 1A) with $Y = \beta_0 + \beta_8 X_8 + \epsilon$. The two fits are as follows.

	Coefficient		Standard Deviation of Coefficients	
Predictor	Rank	Least Squares	Rank	Least Squares
Constant	13.6194	13.6230	0.6929	0.5815
X_8	−0.07992	−0.07983	0.01254	0.01052
	Hodges–Lehmann estimate of tau = 1.061		Least squares s = 0.8901	

The "Hodges–Lehmann estimate of tau" provides an estimate of σ from the robust fit. The two fits are quite similar. An alternative "window estimate" of σ can also be used.

Example 3. Full set of steam data (Appendix 1A) with $Y = \beta_0 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \epsilon$. The two fits are again quite similar.

	Coefficient		Standard Deviation of Coefficients	
Predictor	Rank	Least Squares	Rank	Least Squares
Constant	1.266	1.894	7.287	6.996
X_2	0.5544	0.7054	0.5884	0.5649
X_3	0.358	−1.894	4.319	4.146
X_4	1.4575	1.1342	0.7771	0.7461
X_5	0.1122	0.1188	0.2131	0.2046
X_6	0.14070	0.17935	0.08431	0.08095
X_7	−0.01300	−0.01818	0.02553	0.02451
X_8	−0.07430	−0.07742	0.01728	0.01659
X_9	−0.10952	−0.08585	0.05416	0.05200
X_{10}	−0.4227	−0.3450	0.2195	0.2107
	Hodges–Lehmann estimate of tau = 0.5934		Least squares s = 0.5697	

Other Weights

Other weights, for example, weights derived from the normal distribution, can be used in similar fashion. Quite similar results are obtained in practice. The weights are the normal deviates associated with cumulative normal levels of $100(i - \frac{1}{2})/n$ or, in MINITAB, $100(i - \frac{3}{8})/(n + \frac{1}{4})$, where n is the number of observations and $i = 1, 2, \ldots, n$.

Example 4. Hald data (Appendix 15A), fitting $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Again the two fits are quite similar.

	Coefficient		Standard Deviation of Coefficients	
Predictor	Rank	Least Squares	Rank	Least Squares
Constant	52.222	52.577	3.229	2.286
X_1	1.4631	1.4683	0.1713	0.1213
X_2	0.6678	0.6623	0.0648	0.0459
	Hodges–Lehmann estimate of tau = 3.399		Least squares s = 2.406	

We see throughout these Examples 2–4 that the standard deviation estimates are larger (more conservative) for the rreg fits.

As with many other alternatives to least squares, there is a degree of arbitrariness in the rreg fits in the choice of the weight functions used. In general, we recommend alternative fits mostly to make a comparison with the least squares fit. Large differences seen in the comparison often give useful clues about problems with the data set. Where differences are slight, proceeding with the least squares fit and its subsequent analyses is usually preferable.

Acknowledgment. We are grateful to Monica Hsieh who performed computations for Section 25.6 while pursuing "directed study" credits at the University of Wisconsin–Madison.

25.7. OTHER METHODS

Other robust estimation methods have been suggested. Some of these involve a replacement of one factor of each term in the sum of squares function by a score function, which could be a rank, for example. Currently, such other methods appear to be of limited practical value, and so we do not discuss them.

For a robust version of ridge regression, see Hogg (1979b) and Askin and Montgomery (1980).

25.8. COMMENTS AND OPINIONS

Is it sensible to use robust regression methods? The story here has similarities to that of ridge regression in the following sense. To use a robust estimator blindly is like using a ridge estimator blindly; it may or may not be appropriate. Any specific robust estimator is sensible when it provides (exactly or approximately) maximum likelihood estimation of the parameters under the alternative error assumptions believed to be true, when the assumption $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$ is *not* true. The use of a robust regression method involves three practical difficulties, mentioned earlier:

1. What function $\rho(u)$ should we choose, if we do not know the distribution of the errors in the model? (This amounts to a choice of weight function type, for example, from Figure 25.1.)

2. How should we choose the tuning constants in the $\rho(u)$ we choose? (The values given in Table 25.1 are suggested by past studies. Different values would alter the shapes, but not the basic forms, of the weight functions in Figure 25.1.)

3. How should we choose a robust estimator of the scale factor s in (25.2.6) and (25.2.7)? (Different choices will give different numerical results for the regression coefficients.)

It is often not possible to make a choice of $\rho(u)$ based on factual knowledge about the errors. Usually too little is known. It might make sense simply to look at Figure 25.1 and pick a weight function. In our examples, we picked (b) with the breakpoint at $u = 2$. This would (typically) result in much of the data receiving weight 1, with a few observations receiving lesser weight. Selection (f) is a slight variation of (b). For other choices (c), (d), (e), and (g) only zero residuals receive full weight,

and *all* others are down-weighted. This is more interesting but can also be confusing to assess.

There seems little harm in doing what we showed in our examples, namely, making both a least squares fit *and* a robust fit of some sort, reader's choice. A comparison between the two fits attracts attention to the observations down-weighted by the robust fit (or severely down-weighted if most or all observations receive weights less than 1) and attracts attention to outliers. This knowledge can be combined with, for example, examination of influence statistics like Cook's, and the data can be reassessed, and action taken to choose a good explanatory analysis. Where it is possible to justify a final least squares fit, that is often the best conclusion, simply because we know more about how to interpret the least squares solution.

25.9. REFERENCES

Books

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

Arthanari, T. S., and Y. Dodge (1981). *Mathematical Programming in Statistics.* New York: Wiley.

Birkes, D., and Y. Dodge (1993). *Alternative Methods of Regression.* New York: Wiley.

Bloomfield, P., and W. Steiger (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms.* Boston: Birkhäuser.

Box, G. E. P., and N. R. Draper (1987). *Empirical Model-Building and Response Surfaces.* New York: Wiley.

Dodge, Y. (ed.) (1987a). *Statistical Data Analysis Based on the L_1-Norm and Related Methods.* New York: North-Holland.

Dodge, Y. (ed.) (1987b). Special Issue on Statistical Data Analysis Based on the L_1 Norm and Related Methods. *Computational Statistics & Data Analysis,* vol. 5, pp. 237–450.

Dodge, Y. (ed.) (1992). *L_1-Statistical Analysis and Related Methods.* Amsterdam: North-Holland.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley.

Huber, P. (1981). *Robust Statistics.* New York: Wiley.

Rousseeuw, P. J., and A. M. Leroy (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Staudte, R. G., and S. J. Sheather (1990). *Robust Estimation and Testing.* New York: Wiley.

Articles

Adichie, J. N. (1967). Estimates of regression parameters based on rank test. *Annals of Mathematical Statistics*, **37**, 894–904.

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

Askin, R. G., and D. C. Montgomery (1980). Augmented robust estimators. *Technometrics*, **22**, 333–341.

Barrodale, I., and F. D. K. Roberts (1974). Algorithm 478: solution of an overdetermined system of equations in the l_1-norm. *Communications of the Association for Computing Machinery*, **17**, 319–320.

Beaton, A. E., and J. W. Tukey (1974). The fitting of power series, meaning polynomials, illustrated on band spectroscopic data. *Technometrics*, **16**, 147–185.

Birch, J. B., and D. B. Agard (1993). Robust inferences in regression: a comparative study. *Communications in Statistics, Simulation and Computation*, **22**(1), 217–244.

Davies, P. L. (1987). Asymptotic behaviour of *S*-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics*, **15**, 1269–1292.

Davies, L. (1990). The asymptotics of *S*-estimators in the linear regression model. *Annals of Statistics*, **18**, 1651–1675.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Annals of Statistics*, **20**, 1828–1843.

Dielman, T., and R. Pfaffenberger (1982). LAV (least absolute value) estimation in linear regression: a review. In S. H. Zanakis and J. S. Rustagi (eds.). *Optimization in Statistics*. New York: North-Holland.

Dielman, T., and R. Pfaffenberger (1990). Tests of linear hypotheses and LAV estimation: a Monte Carlo comparison. *Communications in Statistics, Simulation and Computation*, **19**, 1179–1199.

Donoho, D. L., and P. J. Huber (1983). The notion of breakdown point. In P. J. Bickel, K. Doksum, and J. L. Hodges, Jr. (eds.). *A Festschrift for Erich Lehmann* pp. 157–184. Belmont, CA: Wadsworth.

Dutter, R. (1977). Numerical solution of robust regression problems: computational aspects, a comparison. *Journal of Statistical Computing and Simulation*, **5**, 207–238.

Farebrother, R. W. (1988). Algorithm AS 238: a simple recursive procedure for the L_1 norm fitting of a straight line. *Applied Statistics*, **37**, 457–465.

Forsythe, A. B. (1972). Robust estimation of straight line regression coefficients by minimizing pth power deviations. *Technometrics*, **14**, 159–166.

Gentle, J. E., S. C. Narula, and V. A. Sposito (1987). Algorithms for unconstrained L_1 linear regression. In Y. Dodge (ed.). *Statistical Data Analysis Based on the L_1-Norm and Related Methods.* New York: North-Holland.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, **42**, 1887–1896.

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

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Proceedings of the 40th Session of the ISI*, **46**, 375–391.

Hampel, F. R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section*, pp. 59–64. Washington, DC: American Statistical Association.

Hettmansperger, T. P., and S. J. Sheather (1992). A cautionary note on the method of least median squares. *The American Statistician*, **46**, 79–83.

Hill, R. W. (1979). On estimating the covariance matrix of robust regression M-estimates. *Communications in Statistics*, **A8**, 1183–1196.

Hill, R. W., and P. W. Holland (1977). Two robust alternatives to least squares regression. *Journal of the American Statistical Association*, **72**, 828–833.

Hogg, R. V. (1979a). Statistical robustness: one view of its use in applications today. *The American Statistician*, **33**, 108–115.

Hogg, R. V. (1979b). An introduction to robust estimation. In R. L. Launer and G. N. Wilkinson (eds.). *Robustness in Statistics*, pp. 1–18. New York: Academic Press.

Hogg, R. V., and R. H. Randles (1975). Adaptive distribution-free regression methods and their applications. *Technometrics*, **17**, 399–407.

Holland, P. W., and R. E. Welsch (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics A*, **6**, 813–888.

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

Huber, P. J. (1972). Robust statistics: a review. *Annals of Mathematical Statistics*, **43**, 1041–1067.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. *Annals of Statistics*, **1**, 799–821.

Huber, P. J. (1975). Robustness and designs. *A Survey of Statistical Design and Linear Models*. Amsterdam: North-Holland.

Huber, P. J. (1983). Minimax aspects of bounded-influence regression (with discussion). *Journal of the American Statistical Association*, **78**, 66–80.

Jureckova, J., and S. Portnoy (1987). Asymptotics for one-step M-estimators with application to combining efficiency and high breakdown point. *Communications in Statistics, Theory & Methods*, **16**, 2187–2199.

Kelly, G. (1996). Adaptive choice of tuning constant for robust regression estimators. *The Statistician*, **45**, 35–40.

Koenker, R., and G. Bassett (1982). Tests of linear hypothesis and l_1 estimation. *Econometrica*, **50**, 1577–1583.

Krasker, W. S. (1980). Estimation in the linear regression model with disparate data points. *Econometrica*, **48**, 1333–1346.

Krasker, W. S., and R. E. Welsch (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association*, **77**, 595–604.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics*, **17**, 1662–1683.

Lopuhaä, H. P., and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, **19**, 229–248.

Marazzi, A. (1987). Solving bounded influence regression problems with ROBSYS. In: Y. Dodge (ed.). *Statistical Data Analysis Based on the L_1-Norm and Related Methods*. New York: North-Holland.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, **4**, 51–67.

Maronna, R. A., O. Bustos, and V. Yohai (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In T. Gasser and M. Rosenblatt (eds.). *Smoothing Techniques for Curve Estimation*, pp. 91–116. New York: Springer.

Maronna, R. A., and V. J. Yohai (1981). Asymptotic behaviour of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete*, **58**, 7–20.

Maronna, R. A., and V. J. Yohai (1991). The breakdown point of simultaneous general M estimates of regression and scale. *Journal of the American Statistical Association*, **86**, 699–703.

McKean, J. W., and R. M. Schrader (1987). Least absolute errors analysis of variance. In Y. Dodge (ed.). *Statistical Data Analysis Based on the L_1-Norm and Related Methods*. New York: North-Holland.

McKean, J. W., and T. J. Vidmar (1994). A comparison of two rank-based methods for the analysis of linear models. *The American Statistician*, **48**, 220–229.

Morgenthaler, S. (1989). Comment on Yohai and Zamar. *Journal of the American Statistical Association*, **84**, 636.

Ramsay, J. O. (1977). A comparative study of several robust estimates of slope, intercept, and scale in linear regression. *Journal of the American Statistical Association*, **72**, 608–615.

Riedel, M. (1989). On the bias robustness in the location model. I. *Statistics*, **20**, 223–233. II. *Statistics*, **20**, 235–246.

Riedel, M. (1991). Bias-robustness in parametric models generated by groups. *Statistics*, **22**, 559–578.

Rocke, D. M., and D. F. Shanno (1986). The scale problem in robust regression M-estimates. *Journal of Statistical Computing and Simulation*, **24**, 47–69.

Ronchetti, E., and P. J. Rousseeuw (1985). Change-of-variance sensitivities in regression analysis. *Z. Wahrsch. Verw. Gebiete*, **68**, 503–519.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P. J. (1986). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.). *Mathematical Statistics and Application, Vol. B*. Dordrecht: Reidel.

Rousseeuw, P. J., and V. Yohai (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics*, **26**, 256–272. New York: Springer.

Rousseeuw, P. J., and B. C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633–651.

Sadovski, A. N. (1974). Algorithm AS 74: L_1-norm fit of a straight line. *Applied Statistics*, **23**, 244–248.

SAS Institute Inc. (1990). *SAS/STAT User's Guide*, version 6, 4th ed., vol. 2. Cary, NC: SAS Institute.

Schrader, R. M., and T. P. Hettmansperger (1980). Robust analysis of variance based on a likelihood criterion. *Biometrika*, **67**, 93–101.

Schrader, R. M., and J. W. McKean (1987). Small sample properties of least absolute errors analysis of variance. In Y. Dodge (ed.). *Statistical Data Analysis Based on the L_1-Norm and Related Methods*. New York: North-Holland.

Shanno, D. F., and D. M. Rocke (1986). Numerical methods for robust regression: linear models. *SIAM Journal of Scientific and Statistical Computing*, **7**, 86–97.

Simpson, D. G., D. Ruppert, and R. J. Carroll (1992). On one-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, **87**, 439–450.

Stahel, W. A. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. dissertation, Dept. Statistics. ETH, Zürich.

Steele, J. M., and W. L. Steiger (1986). Algorithms and complexity for least median of squares regression. *Discrete Applied Mathematics*, **14**, 93–100.

Stefanski, L. A. (1991). A note on high breakdown estimators. *Statistics and Probability Letters*, **11**, 353–358.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimators for regression. *Annals of Statistics*, **15**, 642–656.

Yohai, V. J., and R. H. Zamar (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406–413.

Welsch, R. E. (1975). Confidence regions for robust regression. *Statistical Computations Section Proceedings of the American Statistical Association*, Washington, DC.

EXERCISES FOR CHAPTER 25

A. Consider any of the data sets in this book and compare the least squares fit with a fit determined by any robust procedure computationally available to you. Do the fits differ much? If yes, what features, or what observations, seem to be causing the difference? Would you be happy using the least squares fit or not?

[Starter suggestions for data: (a) Table 2.1, (b) Appendix 1A, (c) Appendix 15A. One or more predictors can be used.]

B. For the data in Table 2.1, show that the least squares fit is $\hat{Y} = 1.43 + 0.316X$ while the rreg fit with Wilcoxon weights is $\hat{Y} = 1.64 + 0.25X$ and the rreg fit with normal weights is $\hat{Y} = 1.66 + 0.25X$. Plot all three lines. Why is the first fit so different from the other two?