observations; or (4) use a method that is more robust and resistant to outliers. We now turn our attention to various robust regression techniques.

Notes

1. As we shall see in Chapter 6, this is not the case for generalized linear models.

2. With large sample sizes, however, this cutoff is unlikely to identify any observations regardless of whether they deserve attention (Fox 1991).

3. Fox's Influence Plot can be routinely implemented using the influence.plot function in the car package for **R** (see Fox 2002:198 for more details).

4. Partial regression plots should not be confused with the similar partial residual plots. The latter typically are not as effective for assessing influence but tend to be better at distinguishing between monotone and nonmonotone nonlinearity. For more details on the relative merits of the partial regression plots and partial residual plots, see Fox (1997).

4. ROBUST REGRESSION FOR THE LINEAR MODEL

We now explore various robust regression techniques—including those sometimes labeled as resistant regression techniques—in an evolutionary manner, explaining how new methods evolved in response to limitations of existing ones. Several classes of regression will be discussed: *L*-estimators (based on linear combinations of order statistics); *R*-estimators (based on the ranks of the residuals); *M*-estimators (extending from *M*-estimates of location by considering the size of the residuals); *GM*-estimators (or generalized *M*-estimators, which extend *M*-estimators by giving less weight to high influence points as well as to large residual points); *S*-estimators (which minimize a robust *M*-estimate of the residual scale); and *MM*-estimators (which build on both *M*-estimation and *S*-estimation to achieve a high breakdown point with high asymptotic efficiency). Some of these methods should be considered obsolete, but general descriptions are still provided because more recent developments in robust regression build on them. The chapter ends with a discussion of how robust regression can be used as a diagnostic method for identifying problematic cases.

L-Estimators

Any estimator that is computed from a linear combination of order statistics can be classified as an *L*-estimator. The first *L*-estimation procedure, which is somewhat more resistant than OLS, is least absolute values (LAV) regression. Also known as L_1 regression[1] because it minimizes the L_1-norm (i.e., sum of absolute deviations), LAV is the simplest and earliest approach to bounded influence robust regression, predating OLS by about 50 years (Wilcox 2005:451). Least squares regression also fits this definition, and thus it is sometimes referred to as L_2, reflecting that the L_2-norm (i.e., the sum of squared deviations) is minimized. Other well-known L-estimators are the least median of squares and the least trimmed squares estimators.[2]

Least Absolute Values Regression

Least absolute values (LAV) regression is very resistant to observations with unusual y values. Estimates are found by minimizing the sum of the *absolute values of the residuals*

$$\min \sum_{i=1}^{n} |e_i| = \min \sum_{i=1}^{n} \left| y_i - \sum x_{ij}\beta_j \right|. \qquad [4.1]$$

The LAV can be seen as a case of the more general quantile regression. In this case, the objective function to be minimized can be written as

$$\sum_{i=1}^{n} \rho_\alpha(e_i), \qquad [4.2]$$

where

$$\rho_\alpha(e_i) = \begin{cases} \alpha e_i & \text{if } e_i \geq 0 \\ (\alpha - 1)e_i & \text{if } e_i < 0 \end{cases} \qquad [4.3]$$

and α is the quantile being estimated. For general applications of quantile regression, see Koenker and Bassett (1978; see also Koenker and d'Orey 1994; Koenker 2005). For a treatment geared toward social scientists, see Hao and Naiman (2007).[3]

Although LAV is less affected than OLS by unusual y values, it fails to account for leverage (Mosteller and Tukey 1977:366), and thus has a breakdown point of $BDP = 0$. Moreover, LAV estimates have relatively low efficiency. Following the case of the mean, under the assumption that $y \sim N(\mu, \sigma^2)$, the sampling variance of y for OLS is σ^2/n; for LAV it is $\pi/2 = 1.57$ times larger at $\pi\sigma^2/2n$ (in other words, about 64% efficiency). The combination of the low breakdown point and low efficiency makes LAV less attractive than other robust regression methods still to be discussed.

Least Median of Squares Regression

First proposed by Rousseeuw (1984), least median of squares (LMS)[4] replaces the summing of the squared residuals that characterizes OLS with the median of the squared residuals. The estimates are found by

$$\min M\Big(y_{i} - \sum x_{ij}\beta_{j}\Big)^{2} = \min M\Big(e_{i}^{2}\Big), \qquad [4.4]$$

where M denotes the median. The idea is that by replacing the sum with the more robust median, the resulting estimator will be resistant to outliers. Although this result is achieved (it has a breakdown point of $BDP = 0.5$), the LMS estimator has important deficiencies that limit its use. It has at best a relative efficiency of 37% (see Rousseeuw and Croux 1993), and it does not have a well-defined influence function because of its convergence rate of $n^{-1/3}$ (Rousseeuw 1984). Despite these limitations, as we shall see later, LMS estimates can play an important role in the calculation of the much more efficient *MM*-estimators by providing initial estimates of the residuals.

Least Trimmed Squares Regression

Another method developed by Rousseeuw (1984) is least trimmed squares (LTS) regression. Extending from the trimmed mean, LTS regression minimizes the sum of the trimmed squared residuals. The LTS estimator is found by

$$\min \sum_{i=1}^{q} e_{(i)}^{2}, \qquad [4.5]$$

where $q = [n(1-\alpha)+1]$ is the number of observations included in the calculation of the estimator, and α is the proportion of trimming that is performed. Using $q = (n/2)+1$ ensures that the estimator has a breakdown point of $BDP = 0.5$. Although highly resistant, LTS suffers badly in terms of relative efficiency at about 8% (see Stromberg, Hossjer, and Hawkins 2000). Its efficiency is so low that it is not desirable as a stand-alone estimator. Still, the LTS has merit in the role it plays in the calculation of other estimators. For example, the *GM*-estimators proposed by Coakley and Hettmansperger (1993) use LTS to obtain initial estimates of the residuals. LTS residuals can also be used effectively in outlier diagnostic plots, to be discussed later.

R-Estimators

First proposed by Jaeckel (1972), *R*-estimators rely on dispersion measures that are based on the linear combinations of the ordered residuals (i.e., on the rank of the residuals). Let R_i represent the rank of the residuals e_i. *R*-estimators minimize the sum of some score of the ranked residuals

$$\min \sum_{1-1}^{n} a_n(R_i)e_i \qquad [4.6]$$

where $a_n(i)$ is a monotone score function that satisfies

$$\sum_{i=1}^{n} a_n(i) = 0. \qquad [4.7]$$

Many possibilities have been proposed for the score function. The simplest, and perhaps most commonly employed, are the Wilcoxon Scores, which directly find the rank of observations from the median

$$a_n(i) = i - \left(\frac{n+1}{2}\right). \qquad [4.8]$$

Median Scores are a simple adjustment over the Wilcoxon scores,

$$a_n(i) = \sin\left[i - \left(\frac{n+1}{2}\right)\right] \qquad [4.9]$$

Van der Waerden Scores adjust the ranks according to the inverse of the normal probability density function Φ^{-1}:

$$a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right) \qquad [4.10]$$

Finally, Bounded Normal Scores adjust the Van der Waerden Scores by bounding them according to a constant, c:

$$a_n(i) = \min\left\{c, \max\left[\Phi^{-1}\left(\frac{i}{n+1}\right), -c\right]\right\} \qquad [4.11]$$

An advantage of *R*-estimators over some others (such as *M*-estimators, and those extending from them) is that they are scale equivariant. They have some undesirable attributes, however. One problem is that the optimal choice for the score function is unclear. A second problem is that the objective function is invariant with respect to the intercept. If an intercept is not required, this is of no concern—it is simply not estimated. Even if one is needed, it can be calculated manually after fitting the model from the median of the residuals, so this limitation is surmountable. More problematic is the fact that most *R*-estimators have a breakdown point of $BDP = 0$. An exception is the bounded influence *R*-estimator of Naranjo and Hettmensperger (1994), which is also fairly efficient (90%–95%) when the Gauss-Markov assumptions are met. Even for this estimator, however, the breakdown point never reaches more than 0.20. As a result, we leave *R*-estimates behind, proceeding to more robust estimators. (For more extensive details of *R*-estimates, see Huber 2004; Davis and McKean 1993; McKean and Vidmar 1994.)

M-Estimators

First proposed by Huber (1964, 1973, 2004), *M*-estimation for regression is a relatively straightforward extension of *M*-estimation for location. It represents one of the first attempts at a compromise between the efficiency of the least squares estimators and the resistance of the LAV estimators, both of which can be seen as special cases of *M*-estimation. In simple terms, the *M*-estimator minimizes some function of the residuals. As in the case of *M*-estimation of location, the robustness of the estimator is determined by the choice of weight function.

If we assume linearity, homoscedasticity, and uncorrelated errors, the maximum likelihood estimator of β is simply the OLS estimator found by minimizing the sum of squares function

$$\min \sum_{i=1}^{n} \left(y_i - \sum x_{ij}\beta_j \right)^2 = \min \sum_{i=1}^{n} (e_i)^2. \qquad [4.12]$$

Following from *M*-estimation of location, instead of minimizing the sum of squared residuals, a robust regression *M*-estimator minimizes the sum of a less rapidly increasing function of the residuals

$$\min \sum_{i=1}^{n} \rho \left(y_i - \sum x_{ij}\beta_j \right) = \min \sum_{i=1}^{n} \rho(e_i). \qquad [4.13]$$

The solution is not scale equivariant, and thus the residuals must be standardized by a robust estimate of their scale $\hat{\sigma}_e$, which is estimated simultaneously. As in the case of *M*-estimates of location, the median absolute deviation (MAD) is often used. Taking the derivative of Equation 4.13 and solving produces the score function

$$\sum_{i=1}^{n} \Psi \left(y_i - \sum x_{ij}\beta_j \middle/ \hat{\sigma} \right) x_{ik} = \sum_{i=1}^{n} \Psi(e_i/\hat{\sigma}_e)\mathbf{x}_i = 0 \qquad [4.14]$$

with $\Psi = \rho'$. There is now a system of $k+1$ equations, for which Ψ is replaced by appropriate weights that decrease as the size of the residual increases

$$\sum_{i=1}^{n} w_i(e_i/\hat{\sigma}_e)\mathbf{x}_i = 0. \qquad [4.15]$$

Iteratively Reweighted Least Squares

An iterative procedure is necessary to find *M*-estimates for regression. A single step is impossible because the residuals can't be found until the model is fitted, and the estimates can't be found without knowing the residuals. As a result, iteratively reweighted least squares (IRLS) is employed[5]:

1. Setting the iteration counter at $I = 0$, an OLS regression is fitted to the data, finding initial estimates of the regression coefficients $\hat{\beta}^{(o)}$.
2. The residuals are extracted from the preliminary OLS regression, $e_i^{(0)}$, and used to calculate initial estimates for the weights.
3. A weight function is then chosen and applied to the initial OLS residuals to create *preliminary* weights, $w(e_i^{(0)})$.
4. The first iteration, $I = 1$, uses weighted least squares (WLS) to minimize $\sum w_i^{(1)} e_i^2$ and thus obtain $\hat{\beta}^{(1)}$. In matrix form, with $\mathbf{W}$ representing the $n \times n$ diagonal matrix of individual weights, the solution is

$$\hat{\beta}^{(1)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \qquad [4.16]$$

5. The process continues by using the residuals from the initial WLS to calculate new weights, $w_i^{(2)}$.
6. The new weights $w_i^{(2)}$ are used in the next iteration, $I = 2$, of WLS to estimate $\hat{\beta}^{(2)}$.
7. Steps 4–6 are repeated until the estimate of $\hat{\beta}$ stabilizes from the previous iteration.

More generally, at each of the q iterations, the solution is $\hat{\beta}^{(I)} = \left(\mathbf{X}^T\mathbf{W}_q\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_q\mathbf{y}$, where $\underset{(n\times n)}{\mathbf{W}_q} = \text{diag}\left\{w_i^{(I-1)}\right\}$. The iteration process continues until $\hat{\beta}^{(I)} - \hat{\beta}^{(I-1)} \cong 0$. Typically, the solution is considered to have converged when the change in estimates is no more than 0.01% from the previous iteration. We return to IRLS in more detail in Chapter 6 with respect to robust generalized linear models.

M-estimators are defined to be robust against heavy-tailed error distributions and nonconstant error variance—and thus y outliers—but they also implicitly assume that the model matrix $\mathbf{X}$ is measured without error. Under these conditions, *M*-estimates are more efficient than OLS estimates. Under the Gauss-Markov assumptions, however, *M*-estimates are about 95% as efficient as OLS estimates.[6] Moreover, although *M*-estimators are an improvement over OLS in terms of resistance and robustness to regression outliers (i.e., unusual y values given their xs), like LAV estimators, they are not completely immune to unusual observations because they do not consider leverage. Recall that *M*-estimates of location are highly robust, having a bounded influence function and a breakdown point of $BDP = 0.5$. *M*-estimates for regression share these attributes for y but not for the xs, resulting in a breakdown point of $BDP = 0$. In other words, in some situations they perform no better than OLS (see Rousseeuw and Leroy 1987). As we shall see later, these estimators are still important because of the role they play in computing other, more robust estimates. Because these newer estimates perform much better, they should generally be preferred over the original *M*-estimation.

GM-Estimators

The *M*-estimator has unbounded influence because it fails to account for leverage (Hampel et al. 1986). In response to this problem, bounded influence Generalized *M*-estimators (*GM*-estimators) have been proposed. The goal was to create weights that consider both vertical outliers *and* leverage. Outliers are dealt with using a standard *M*-estimator, and leverage points are typically down-weighted according to their hat value. The general *GM* class of estimators is defined by

$$\sum_{i=1}^{n} w_i(\mathbf{x}_i)\Psi\left\{\frac{e_i}{v(\mathbf{x}_i)\hat{\sigma}_e}\right\}\mathbf{x}_i = 0, \qquad [4.17]$$

where Ψ is the score function (as in the case of *M*-estimation, this is typically the Huber or biweight function), and the weights w_i and v_i initially depend on the model matrix $\mathbf{X}$ from an initial OLS regression fitted to the data but are updated iteratively.

The first *GM*-estimator proposed by Mallows (see Krasker and Welsch 1982) includes only the w_i weights—that is, $v_i(\mathbf{x}_i) = 1$ in Equation 4.17. The w_i are calculated from the hat values. Because hat values range from 0 to 1, a weight of $w_i = \sqrt{1 - h_i}$ ensures that observations with high leverage receive less weight than observations with small leverage (i.e., if $h_i > h_j$, $u_i < u_j$). Although this strategy seems sensible at first, it is problematic because even "good" leverage points that fall in line with the pattern in the bulk of the data are down-weighted, resulting in a loss of efficiency.

Schweppe's solution (introduced in Handschin et al. 1975) adjusts the leverage weights according to the size of the residual e_i. In order to achieve this result, the w_i weights are defined in the same way as for Mallows, $w_i = \sqrt{1 - h_i}$, but now $v_i(\mathbf{x}_i) = w_i$ (see Chave and Thomson 2003). Although the breakdown point for Schweppe's estimators is better than for regular *M*-estimators that don't consider leverage, Maronna, Butos, and Yohai (1979) show that it is never higher than $1/(p+1)$, where p is the number of parameters estimated by the model. In other words, as dimensionality increases, the breakdown point gets closer to $BDP = 0$. This is especially problematic because as the number of variables in the model increases, detection of influential cases also becomes increasingly more difficult. Moreover, because they down-weight according to x values without considering how the corresponding y values fit with the pattern of the bulk of the data, efficiency is still hindered (see Krasker and Welsch 1982). Other evidence also suggests that the Schweppe estimator is not consistent when the errors are asymmetric (Carroll and Welsh 1988), meaning that they are ineffective for the more common problem of outliers in one of the tails, the main concern of the present book.

In an attempt to overcome these problems, other *GM*-estimation procedures completely remove severe outliers and then use *M*-estimation on the remaining "good" observations (Coakley and Hettmansperger 1993; Chave and Thomson 2003). Perhaps the most notable of these is Coakley and Hettmansperger's (1993) *Schweppe one-step estimator* (S1S), which extends from the original Schweppe estimator. The advantage of this estimator over the original is that the leverage weights consider where observations fit with the bulk of the data. In other words, it considers whether the observations are "good" or "bad" leverage points, giving less weight to the latter. This results in 95% efficiency relative to OLS estimators under the Gauss-Markov assumptions.

The S1S estimator takes initial estimates of the residuals and the scale of the residuals from a regression with a high breakdown point rather than from an OLS regression, as is the case with the *GM*-estimators developed before it. Using Rousseeuw's LTS estimator for the initial estimates gives a breakdown point of $BDP = 0.5$. The method is also different from the Mallows and Schweppe estimations in that once the initial estimates from the LTS regression are included, final *M*-estimates are calculated in a single step (hence the name "one-step") rather than iteratively. Although S1S estimators are more efficient than other *GS*-estimators, and in fact quite comparable to OLS estimators under normality and large sample sizes, simulation studies suggest that their efficiency is very low when *n* is small (see Wilcox 2005:438-440).

S-Estimators

In response to the low breakdown point of *M*-estimators, Hampel (1975) suggested considering the scale of the residuals. Following this idea, Rousseeuw and Yohai (1984; see also Rousseeuw and Leroy 1987) proposed *S*-estimates. *S*-estimates are the solution that finds the smallest possible dispersion of the residuals

$$\min \hat{\sigma}\big(e_1(\hat{\beta}), \ldots e_n(\hat{\beta})\big). \qquad [4.18]$$

The parallel with OLS, which minimizes the variance of the residuals, should be obvious. Hence, OLS can be seen as a special, less robust case of *S*-estimation. Rather than minimize the variance of the residuals, robust *S*-estimation minimizes a robust *M*-estimate of the residual scale

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}_e}\right) = b, \qquad [4.19]$$

where b is a constant defined as $b = E_\Phi[\rho(e)]$ and Φ represents the standard normal distribution. Differentiating Equation 4.19 and solving results in

$$\frac{1}{n}\sum_{i=1}^{n} \Psi\left(\frac{e_i}{\hat{\sigma}_e}\right) = b, \qquad [4.20]$$

where Ψ is replaced with an appropriate weight function. As with most *M*-estimation procedures, either the Huber weight function or the biweight function is usually employed. Although *S*-estimates have a breakdown point of $BDP = 0.5$, it comes at the cost of very low efficiency (approximately 30%) relative to OLS (Croux, Rousseeuw, and Hossjer 1994).

Generalized *S*-Estimators

Croux et al. (1994) propose *Generalized S-estimates* (*GS*-estimates) in an attempt to overcome the low efficiency of the original *S*-estimators. These estimators are computed by finding a *GM*-estimator of the scale of the residuals. A special case of the *GS*-estimator is the *least quartile difference* *estimator* (LQD), the parallel of which is using the interquartile range to estimate the scale of a variable. The LQD estimator is defined by

$$\min Q_n(e_1, \ldots, e_n), \qquad [4.21]$$

where

$$Q_n = \left\{ \left| e_i - e_j \right|; i < j \right\}_{\binom{h_p}{2} : \binom{n}{2}} \qquad [4.22]$$

and

$$h_p = \frac{n + p + 1}{2} \qquad [4.23]$$

and p is the number of parameters in the model. Put more simply, this means that Q_n is the $\binom{h_p}{2}$th order statistic among the $\binom{n}{2}$ elements of the set $\left\{ \left| e_i - e_j; i < j \right| \right\}$. Although these estimators are more efficient than *S*-estimators, they have a "slightly increased worst-case bias" (Croux et al. 1994:1271).

Yohai and Zamar's (1988) τ estimates are also defined by the minimization of an estimate for the scale of the residuals, but the weights are adaptive depending on the underlying error distribution. This results in a high breakdown point and a high efficiency estimate of the scale of the errors. Nevertheless, points with high leverage are not considered, so the estimator's efficiency is still hindered. Ferretti et al. (1999) tried to overcome this limitation with *generalized τ estimates*, which use weights that consider observations with high leverage in much the same way as *GM*-estimates extend from *M*-estimates. The method achieves a high breakdown point (as high as 0.5) and higher efficiency (though still only about 75%) than other *GS*-estimates. Still, 75% efficiency is low compared to many other estimators, limiting the use of *S*-estimators as stand-alone estimators. On the other hand, because they are highly resistant to outliers, *S*-estimators play an important role in calculating *MM*-estimates, which are far more efficient.

MM-Estimators

First proposed by Yohai (1987), *MM*-estimators have become increasingly popular and are perhaps now the most commonly employed robust regression technique. They combine a high breakdown point (50%) with good efficiency (approximately 95% relative to OLS under the Gauss-Markov assumptions). The "*MM*" in the name refers to the fact that more than one *M*-estimation procedure is used to calculate the final estimates. Following from the *M*-estimation case, iteratively reweighted least squares (IRLS) is employed to find estimates. The procedure is as follows:

1. Initial estimates of the coefficients $\hat{\beta}^{(1)}$ and corresponding residuals $e_i^{(1)}$ are taken from a highly resistant regression (i.e., a regression with a breakdown point of 50%). Although the estimator must be consistent, it is not necessary that it be efficient. As a result, *S*-estimation with Huber or bisquare weights (which can be seen as a form of *M*-estimation) is typically employed at this stage.[7]
2. The residuals $e_i^{(1)}$ from the initial estimation at Stage 1 are used to compute an *M*-estimation of the scale of the residuals, $\hat{\sigma}_e$.
3. The initial estimates of the residuals $e_i^{(1)}$ from Stage 1 and of the residual scale $\hat{\sigma}_e$ from Stage 2 are used in the first iteration of weighted least squares to determine the *M*-estimates of the regression coefficients

   $$\sum_{i=1}^{n} w_i\left(e_i^{(1)}\Big/\hat{\sigma}_e\right)\mathbf{x}_i = 0, \qquad [4.24]$$

   where the w_i are typically Huber or bisquare weights.

4. New weights are calculated, $w_i^{(2)}$, using the residuals from the initial WLS (Step 3).
5. Keeping constant the measure of the scale of the residuals from Step 2, Steps 3 and 4 are continually reiterated until convergence.

Comparing the Various Estimators

Table 4.1 summarizes some of the robustness attributes of most of the estimators we have discussed. Reported are the breakdown point, whether or not the estimator has a bounded influence function, and the approximate asymptotic efficiency of the estimator relative to the OLS estimator. Immediately obvious is the comparatively low breakdown point for the LAV and *M*-estimators, which, depending on how the data are configured, can sometimes perform no better than OLS estimators. A single discrepant observation can render these estimates useless. The bounded influence *R*-estimators don't do much better, having a breakdown point of less than $BDP = 0.2$. These methods, at least on their own, should be ignored in favor of others.

TABLE 4.1
Robustness Attributes of Various Regression Estimators

Estimator	*Breakdown Point*	*Bounded Influence*	*Asymptotic Efficiency*
OLS	0	No	100
LAV	0	Yes	64
LMS	.5	Yes	37
LTS	.5	Yes	8
LTM	.5	Yes	66
Bounded *R*-estimates	$< .2$	Yes	90
M-estimates (Huber, biweight)	0	No	95
GM-estimates (Mallows, Schweppe)	$1/(p+1)$	Yes	95
GM-estimates (S1S)	.5	Yes	95
S-estimates	.5	Yes	33
GS-estimates	.5	Yes	67
Generalized estimates	.5	Yes	75
MM-estimates	.5	Yes	95

We should also be cautious of estimators with low efficiency, such as the LMS, LTS, LTM, and *S*-estimates. If the goal is only to ensure resistance, and not to make inferences about a population, then these estimates may be appropriate. On the other hand, one should not use them without good knowledge of the nature of the unusual observations. Using them blindly could result in less efficient estimates than otherwise possible. For example, if the errors are normally distributed, it would be better to use OLS estimators.

Despite its low breakdown point, the efficiency of the *M*-estimates is a favorable attribute. When used in combination with more resistant estimators, new estimators result that are both highly resistant to outliers and highly efficient. *M*-estimation using the residuals from an initial highly resistant LTS fit, for example, leads to the S1S *GM*-estimator, which is highly resistant to both residual outliers and high leverage observations, and yet maintains an efficiency of about 95% relative to the OLS estimator. Computing *M*-estimates from the residuals from an LMS or *S*-estimation leads to the similarly efficient and robust *MM*-estimators.

Example 4.1: Simulated Data

We now return to the simulated data first introduced in Chapter 3 that include various types of outliers. We explore six different regression estimates for each of the three ‘‘contaminated’’ data sets: an OLS estimate, an LAV estimate, an *M*-estimate (using Huber weights), a *GM*-estimate (specifically a Coakley-Hettmansperger estimator), an *S*-estimate, and an *MM*-estimate. The fitted lines for the estimates for each of the data sets are shown in Figure 4.1, with the ‘‘contaminating’’ observation identified.

Starting with the vertical outlier scenario, the substantive conclusions, at least in terms of the slope coefficients, are nearly identical regardless of which regression method is employed. In fact, aside from the OLS line, which stands apart only slightly in terms of its intercept, it is impossible to distinguish the lines for the other methods from each other. Moreover, although the OLS intercept is slightly smaller than the intercept for the others—indicating that the line is pulled toward the outlier—it is not so dissimilar from the others that it is problematic. The estimates for the ‘‘good’’ leverage point (B) are even more similar to each other. Aside from the LAV line, which falls just slightly above the others, all of the other regression lines are directly on top of each other. For the case of the ‘‘bad’’ leverage point (C), the various estimates differ to a greater degree, although the most marked differences are with respect to contrast between the OLS estimate and the others. As we saw in Chapter 3, the OLS regression line has been pulled toward the influential observation. None of the robust regression estimates is substantially influenced by the outlier, however. It is clear, then, that a more robust method should be favored in this last scenario. But what about the other two scenarios, for which very little difference was found between the estimates?

To answer this question, we turn to the distribution of the residuals to see whether the precision of the OLS estimates might be hindered. Recall that standard errors for the OLS estimates are smallest when there is less spread in the residuals. Figure 4.2, which shows the distribution of the residuals, suggests that the vertical outlier causes problems for the standard errors of the OLS estimators (see plot A). On the other hand, the residuals are very well behaved despite the addition of the ‘‘good’’ leverage point (plot B). The information in Figures 4.1 and 4.2 taken together suggest that OLS is only suitable for the data with the good leverage point. Given that all of the robust estimators tell the same general story, an efficient estimator like the *MM*-estimator would be a good choice for the other scenarios.

Example 4.2: Multiple Regression Predicting Public Opinion

We now return to the cross-national public opinion data, continuing to focus on only the new democracies. Earlier, we used OLS to fit a model predicting public opinion from per capita GDP and the Gini coefficient. Diagnostics and preliminary analyses suggested that the model performed better if the Czech Republic and Slovakia were omitted (see Table 3.2). Recall that the OLS, including all observations, showed a statistically significant

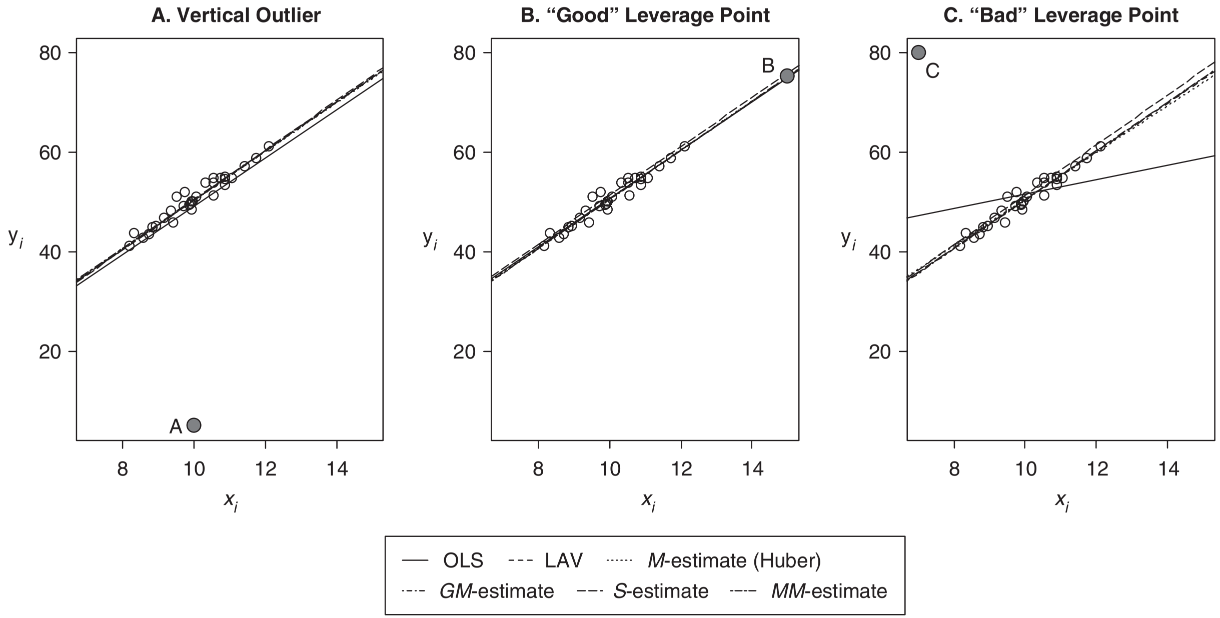

Figure 4.1 Various Regression Estimates for the Contrived Data Including Three Different Types of Outliers

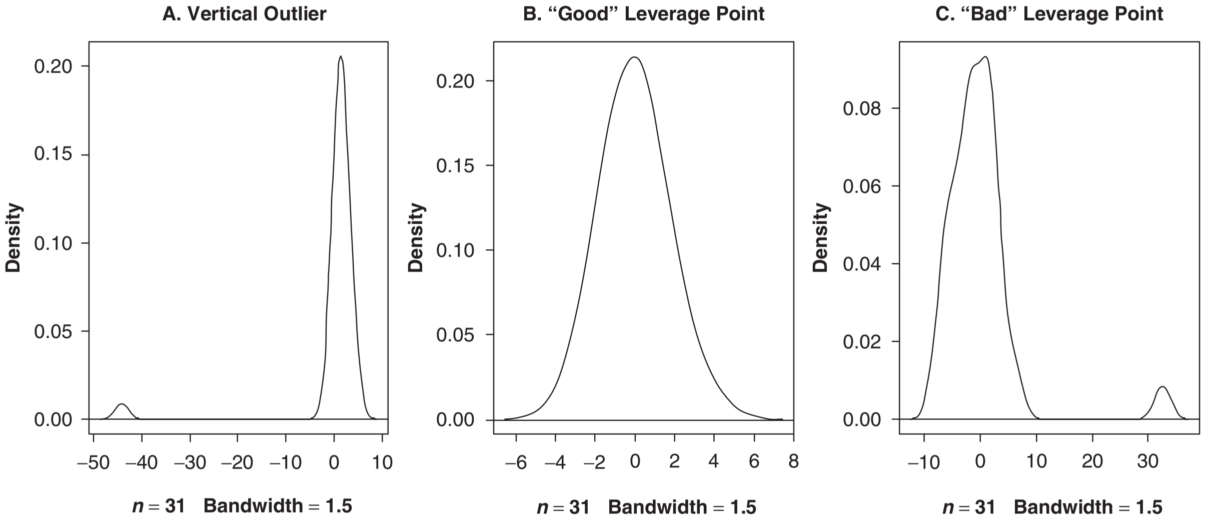

Figure 4.2 Density Estimates of the Residuals for the OLS Regression Fitted to the Three ‘‘Contaminated’’ Data Sets

TABLE 4.2
Robust Regression Models Fitted to the Public Opinion Data, New Democracies

	LAV Regression	*M-Estimation (Huber)*	*M-Estimation (Biweight)*	*MM-Estimation*	*Generalized M-Estimation (Coakley-Hettmansperger)*
Intercept	–0.079	–0.063	–0.091	–0.097	0.939
Gini	0.0045	0.0039	0.0049	0.0051	0.0041
Per Capita GDP/1000	0.0059	0.0089	0.0052	0.0057	0.0065
n	26	26	26	26	26

positive effect of per capita GDP ($\hat{\beta} = 0.0175$) and a statistically insignificant effect for the Gini coefficient ($\hat{\beta} = 0.00074$). After the two outliers were removed, the coefficient for per capita GDP fell to about one third the size and was no longer statistically significant ($\hat{\beta} = 0.0063$), whereas the slope for the Gini coefficient became more than seven times as large and statistically significant ($\hat{\beta} = .00527$).

Table 4.2 gives the estimates from several robust regressions fitted to the same data. Although there are small differences between them, the *M*-estimators, *MM*-estimator, and *GM*-estimator tell a similar story regarding the effects of per capita GDP and the Gini coefficient. All of these methods give results similar to the OLS regression that omits the two outliers. The LAV regression also does a good job uncovering the relationship between the Gini coefficient and public opinion but gives a much smaller estimate for the effect of per capita GDP, which, in any event, does not have a statistically significant effect for the OLS regression without outliers. In summary, the robust regression methods did a much better job of handling the influential cases than did the ordinary least squares regression.

Diagnostics Revisited: Robust Regression-Related Methods for Detecting Outliers

The discussion above shows the merit of robust regression in limiting the impact of unusual observations. Although it is certainly sensible to see it as the final method to report, it can also be used in a preliminary manner as a diagnostic tool (see Atkinson and Riani 2000). In this respect, it can be a good complement to the traditional methods for detecting unusual cases that were discussed in Chapter 3.

A criticism of common measures of influence, such as Cook's D, is that they are not robust. Their calculation is based on the sample mean and covariance matrix, meaning that they can often miss outliers (see Rousseeuw and van Zomeren 1990). More specifically, Cook's D is prone to a "masking effect," where a group of influential points can mask the impact of each other. We have already seen that partial regression plots can be helpful for overcoming the masking effect with respect to individual coefficients. Information from the weights and residuals from robust regression can help combat the masking problem when assessing overall influence on the regression.

Index Plots of the Weights From the Final IWLS Fit

A straightforward way to use robust regression as a diagnostic tool involves the weights from the final IWLS fit. It is important to remember, however, that the meaning of the weights differs according to the model. Different methods could give quite different weights to observations depending on the type of unusualness. For *M*-estimation, the only thing that we can say about the weights is that they indicate the size of the residuals in an OLS fit, and thus whether or not the observation is a vertical outlier. Examined alone, they tell us nothing about leverage, and thus influence, because *M*-estimates do not take these into account when assigning the weights. *GM*-estimation, on the other hand, down-weights observations according to both the size of their residual in an OLS fit and their leverage, although an examination of the weights will not allow us to distinguish between the two aspects. Weights from the *MM*-estimate can provide a good indication of overall influence because of their highly resistant initial step.

To evaluate how the weights from various robust regressions perform, it is instructive to compare them to the Cook's distances and other outlier detection measures from an OLS regression fitted to the same data. Table 4.3 contains this information. We found in Chapter 3 that Cook's distances identified the Czech Republic and Slovakia as influential observations. All of the robust regression also detected the two discrepant observations, giving them comparatively low weight. In other words, the weights indicate the level of unusualness—the smaller the weight, the more unusual is the observation.

The evidence from Table 4.3 motivates the idea of plotting robust regression weights in an index plot in the same manner as is typically done with Cook's D. This has been done in Figure 4.3. Although all three robust regression methods identified the two most problematic observations, the *GM*-estimation gave nine other observations a weight less than .5, whereas neither of the other methods gave any other observation a weight less

TABLE 4.3
Diagnostic Information From OLS and Robust Regressions

	OLS Diagnostic Statistics			*Final Weights From Robust Regressions*			
Country	*Cook's D*	*Hat Value*	*Studentized Residual*	*M-Estimate (Huber)*	*M-Estimate (Bisquare)*	*MM-Estimate*	*GM-Estimate*
Armenia	0.0030	0.10	−0.27	1	0.81	0.87	0.35
Azerbaijan	0.0011	0.10	−0.17	1	0.97	0.98	0.86
Bangladesh	0.0012	0.18	−1.13	1	1	1	1
Belarus	0.0096	0.07	−1.63	1	1	1	1
Brazil	0.012	0.26	0.36	1	1	1	1
Bulgaria	0.0011	0.07	0.21	1	0.82	0.88	0.38
Chile	0.1350	0.24	1.14	0.72	0.55	0.72	0.18
China	0.0005	0.073	0.13	1	1	1	1
Croatia	0.0073	0.063	−0.56	1	1	1	1
Czech Republic	0.3629	0.17	2.60	0.22	0	0	0.62
Estonia	0.0155	0.04	−1.00	1	0.85	0.89	0.40
Georgia	0.0011	0.07	−0.21	1	0.98	0.99	1
Hungary	0.0273	0.098	−0.87	1	1	1	1
Latvia	0.0093	0.07	−0.59	1	1	1	1
Lithuania	0.0040	0.04	−0.51	1	0.99	1	1
Mexico	0.0003	0.17	0.06	1	1	1	1

	OLS Diagnostic Statistics			*Final Weights From Robust Regressions*			
Country	*Cook's D*	*Hat Value*	*Studentized Residual*	*M-Estimate (Huber)*	*M-Estimate (Bisquare)*	*MM-Estimate*	*GM-Estimate*
Moldova	0.0038	0.112	0.295	1	0.96	0.97	0.84
Nigeria	0.0504	0.14	0.95	1	0.89	0.93	0.47
Romania	0.0004	0.08	−0.11	1	0.91	0.94	0.55
Russia	0.0189	0.09	−0.76	0.87	0.68	0.77	0.27
Slovakia	0.6990	0.17	4.32	0.17	0	0	0.05
Slovenia	0.2691	0.24	−1.66	1	0.99	0.99	1
Taiwan	0.1296	0.15	−1.50	1	0.96	0.97	0.80
Turkey	0.0011	0.05	0.26	1	0.89	0.93	0.46
Ukraine	0.0024	0.1	−0.26	1	0.83	0.88	0.38
Uruguay	0.0059	0.07	0.46	0.95	0.65	0.78	0.24

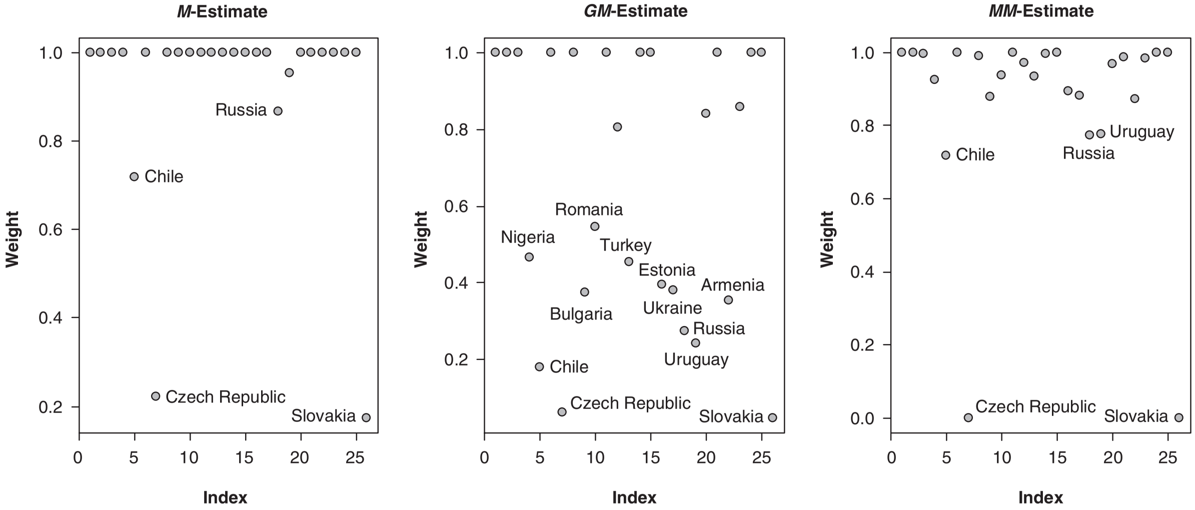

Figure 4.3 Index Plots of Final Weight From the IWLS Fit for Various Robust Regression Estimates

than .7. As said above, the uniqueness of the *GM*-estimate results because it considers the size of the residual *and* leverage.

RR-Plots ("Residual-Residual" Plots)

According to Rousseeuw and van Zomeren (1990:637), robust regression residuals are much better than OLS residuals for diagnosing outliers because the OLS regression "tries to produce normal-looking residuals even when the data themselves are not normal." With this in mind, Tukey (1991) proposed the RR-plot ("residual-residual" plot), which calls for a scatterplot matrix that includes plots of the residuals from an OLS fit against the residuals from several different robust regressions. If the OLS assumptions hold perfectly, there will be a perfect positive relationship, with a slope equal to 1 (called the "identity line"), between the OLS residuals and the residuals from any robust regression. Let the *i*th residual from the *j*th regression fit $\hat{\beta}_j$ be $e_{ij} = y_i - \mathbf{x}_t^T \hat{\beta}_j$, then

$$\begin{aligned} \|e_i(\hat{\beta}_1) - e_i(\hat{\beta}_2)\| = \|\hat{y}_i(\hat{\beta}_1) - \hat{y}_i(\hat{\beta}_2)\| = \|\mathbf{x}_i^T(\hat{\beta}_1 - \hat{\beta}_2)\| \\ \leq \|\mathbf{x}_i\|\left(\|\hat{\beta}_1 - \beta\| + \|\hat{\beta}_2 - \beta\|\right). \end{aligned} \qquad [4.25]$$

This implies that as n approaches ∞, the scatter around the identity line will get tighter and tighter if the regression assumptions are met. If there are outliers, the slope will be a value other than 1 because the OLS regression does not resist them whereas the robust regression does.

RR-plots for the public opinion data are shown in Figure 4.4. The broken line is the identity line; the solid line shows the regression of the residuals from the method on the vertical axis on the residuals from the method on the horizontal axis. The plots in the first column are of most interest because they show the regression of the OLS residuals on the residuals from various robust regression methods. The fact that the two lines are far apart from each other in all of these plots indicates that the OLS estimates were highly influenced by the outliers; the Czech Republic and Slovakia have much smaller residuals for the OLS regression, indicating that they are quite influential. Turning to the other plots, we notice that the residuals from the various robust regressions are very similar to each other, especially with respect to the *MM*-estimates and *GM*-estimates, which are nearly identical.

Robust Distances

We can also consider diagnostic methods that pertain only to a robust regression. For example, Rousseeuw and van Zomeren (1990) claim that a plot of the robust residuals against robust distances, the latter being based on the Mahalanobis distance but defined by a robust covariance matrix, is

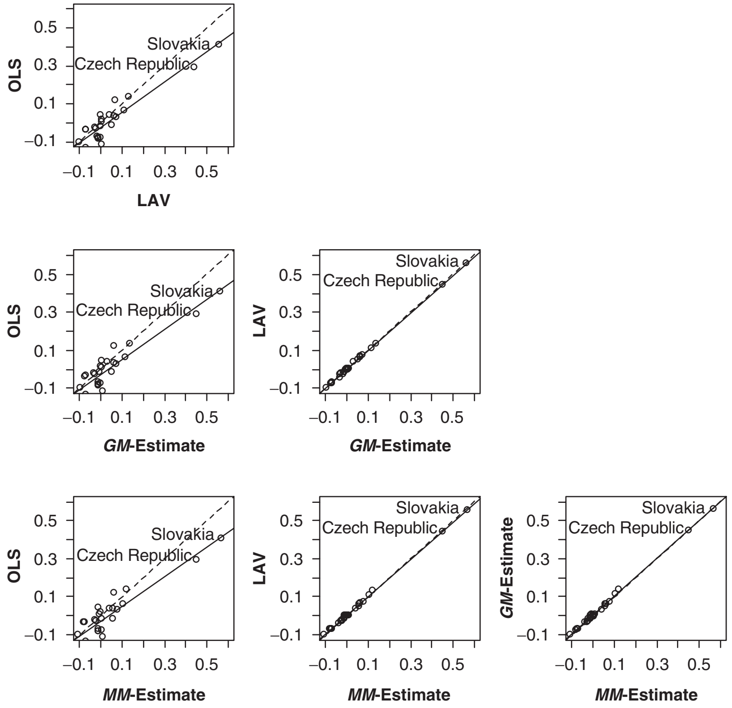

Figure 4.4 RR-Plots for the Regression of Public Opinion Regressed on Per Capita GDP and the Gini Coefficient, New Democracies

able to detect multiple outliers better than traditional methods (see also Cook and Hawkins 1990; Ruppert and Simpson 1990; and Kempthorne and Mendel 1990 for debate about this topic). Efficiency is not a concern for these diagnostics, so the residuals from the highly resistant LMS or LTS regressions are often employed.

The Mahalanobis distance measures how far an observation x_i is from the center of the cloud of points defined by the data set **X.** It is defined by

$$\text{MD}_i = \sqrt{\left(\mathbf{x}_i - \bar{\mathbf{x}})\text{cov}(\mathbf{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T\right.}, \qquad [4.26]$$

where $\bar{\mathbf{x}}$ is the centroid of $\mathbf{X}$ and $\text{cov}(\mathbf{X})$ is the sample covariance matrix. Outliers can influence the mean and covariance matrix, and thus they will not necessarily be detected by the MD_i. As a result, Rousseeuw and van Zomeren's (1990) robust distances RD_i are defined by replacing $\text{cov}(\mathbf{X})$ and $\bar{\mathbf{x}}$ with

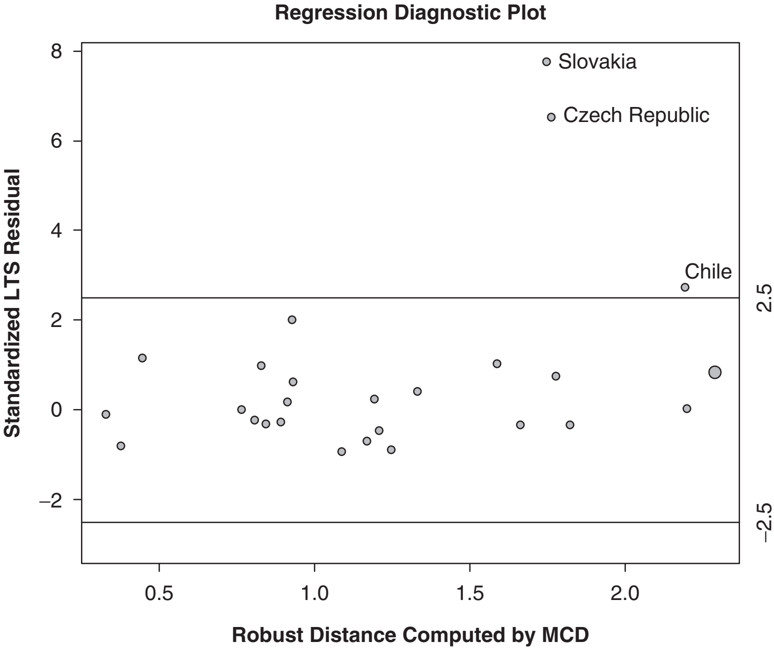

Figure 4.5 Plot of Robust Residuals (From LTS Fit) Against Robust Distances

the more robust center and covariance matrix from the minimum volume ellipsoid estimator (see Rousseeuw 1985 for more details). Usual practice is to identify standardized robust residuals as problematic if they are $e' \geq |2.5|$. Similarly, robust distances are identified as having high leverage if $\mathrm{RD}_i > 0.975$ percent point of the chi-squared distribution with degrees of freedom equal to the number of parameters estimated in the model.

Rousseeuw and van Zomeren's regression diagnostic plot for the public opinion data is shown in Figure 4.5. Plotted against the robust distances are the standardized residuals from a LTS regression. Although the robust distances indicate that no cases have unusually high leverage, the robust residuals suggest that three cases are outliers. Following in line with the rest of the analyses that we have done so far, Slovakia and the Czech Republic are two of these. The third observation, which is just barely past the rule of thumb cutoff, is Chile.

As well as the methods discussed above, the traditional diagnostic plots for identifying outliers (discussed in Chapter 3) can be extended to robust regression models. Given that they are generally interpreted in the same way as for the OLS fit, they're not discussed here. For more information on these diagnostics, see McKean and Sheather (2000). Other techniques related to robust regression can also be seen in Fung (1999) and Pena and Yohai (1999).

Notes

1. Other names for LAV regression are least absolute deviations (LAD) regression and minimum sum of absolute errors (MSAE) regression (Birkes and Dodge 1993).

2. Related methods not discussed in this book because of their limited use are the least-trimmed median estimators and the least-trimmed difference estimators. Both of these have breakdown points of $BDP = 0.5$, but their relative efficiency is less than 67%. For more information, see Croux et al. (1994) and Stromberg et al. (2000).

3. These estimators are sometimes referred to as *trimmed-mean* estimators. They can also be modified so that they have a bounded influence function (see De Jongh, De Wet, and Welsh 1988).

4. The LMS estimator should not be confused with Siegel's (1982) repeated median (*RM*). Although proposed as a robust estimator quite early, the *RM* estimator has the severe limitation of not being affine-regression equivariant (i.e., coefficient estimates do not behave as expected when the predictors are rescaled or combined in linear ways) for high-dimensional problems. It will not be discussed any further because of this limitation.

5. IRLS is also often referred to as *iterative weighted least squares* (IWLS).

6. This assumes tuning constants of $c = 1.345$ for Huber weights and $c = 4.685$ for biweights. See the earlier discussion of *M*-estimation of location for more details.

7. Other methods, such as LMS estimation (Rousseeuw 1984) and RM estimation (Siegel 1982), have also been proposed for the initial stage.

5. STANDARD ERRORS FOR ROBUST REGRESSION

Analytical standard errors are easily calculated for some, but not all, types of robust regression.[1] Nonetheless, even when analytical standard errors can be calculated, they are not reliable for small samples. As a result, it is often desirable to use bootstrapping to calculate standard errors. As a result, this chapter starts with a brief discussion of asymptotic standard errors, and then continues by exploring various types of bootstrapped standard errors and confidence intervals.