Modern Methods of Data Analysis

edited by

John Fox
J. Scott Long

SAGE PUBLICATIONS
The International Professional Publishers
Newbury Park London New Delhi

7

A PRIMER ON ROBUST REGRESSION

Richard A. Berk

INTRODUCTION

For more than two decades, least squares estimation has dominated multivariate analyses in the social sciences. Much like cross-tabulation for an earlier generation, analysis of variance, analysis of covariance, and multiple regression, often extended to multiple equation applications, have become basic tools of the trade. With the more recent interest in latent variables, maximum likelihood estimation procedures also have become popular, but when the normal distribution is invoked, the least squares criterion is still effectively in place. Indeed, maximum likelihood estimation for the full set of generalized linear models may be properly undertaken with iteratively reweighted least squares. These generalized linear models include not just the conventional linear regression, but such popular techniques as logistic regression, probit analysis, and log-linear techniques for contingency tables (McCullagh & Nelder, 1983). In short, the vast majority of estimation procedures currently used in sociology rely on at least the equivalent of a least squares "fit."

It is widely recognized that estimators associated with the least squares principle are especially sensitive to larger residuals. In effect, the estimates produced take particular account of larger "errors." If by a "good fit" one means responding to larger residuals, all may be well. However, if by a good fit one means protecting against larger residuals,

AUTHOR'S NOTE: Thanks go to Jan De Leeuw for helpful comments on an earlier version of this chapter and to Alice Hoffman for help in constructing the difficult tables.

fitting by least squares will often give misleading answers. The mean, for example, rests on a least squares fit and is well known to be misleading for asymmetric distributions with long tails.

In this chapter, I consider M-estimators for regression analysis, which, as one kind of robust location estimator, do not depend on the least squares principle. M-estimators can minimize many different functions of the residuals, not just the sum of their squared values. As a result, M-estimators can weight observations in a variety of ways. Other robust estimators of location share these characteristics, but M-estimators for *robust regression* have excellent statistical properties, may be easily modified for particular problems, and are relatively easy to compute (Hampel et al., 1986; Li, 1985; Wu, 1985; Chapter 2, this volume). Whereas by these criteria M-estimators probably dominate the field,[1] I will not be discussing the full variety of robust regression procedures, many of which have considerable merit. Nor will I tackle in any depth the more general issues associated with robust statistics.[2] Both are well beyond the scope of a single introductory chapter.

In the next section, I provide a broad overview of the issues that motivate the material to follow. In the third section, I briefly consider the formal definition of M-estimators and summarize how M-estimators perform. It cannot be overemphasized that my exposition will be no more than an introduction and that interested readers should consult the references cited. Then, in the fourth and fifth sections, I undertake some data analyses showing how robust regression may be applied. The data in the fifth section are particularly instructive because there are far too few observations to capitalize on the statistical convenience of asymptotic (i.e., large sample) distributions. Finally, in the last section, I extract some general lessons.

SOME BACKGROUND

Problems with quadratic objective functions are well documented in the statistical literature, and excellent discussions can now be found in a few elementary texts (e.g., Mosteller, Fienberg, & Roarke, 1985, Mosteller & Tukey, 1977).[3] Briefly, there are two generic concerns: diagnosis and cure. Under diagnosis falls a very rich and useful tradition in statistics, including the detection of anomalous observations and determinations of the impact of those observations on one's results (Barnett & Lewis, 1978; Belsley, Kuh, & Welch, 1980; Cook &

Weisberg, 1982; Chapters 5 and 6, this volume). Whereas these diagnostic procedures have been developed from a variety of perspectives and in reaction to a number of particular problems, many speak effectively to the ways in which quadratic objective functions may be inappropriate for certain kinds of data.

Under the heading of cure, there have been two related strategies. On the one hand, there are situations in which observations that are clearly anomalous (*outliers*) result from known measurement errors or known flaws in the execution of a research design. It is then possible either to correct the troublesome data or to delete them. For example, perhaps one's problems derive from the transposition of digits during coding or from the inadvertent aggregation in only some units (e.g., school districts) being studied. In both cases, if the errors cannot be corrected, one may choose to discard the offending observations. I will not consider such options in this chapter, but suffice it to say that one must have a convincing explanation for how the errors were introduced (Barnett & Lewis, 1978; chap. 2; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; pp. 56-71; Chapter 6, this volume).

On the other hand, there will often be times when it is not clear which observations are anomalous. They may not appear to be dramatically different from the rest of the data, and/or they are not readily explained by any known error in measurement or data collection. Indeed, it is all too easy to forget that improbable events occur; what looks to be a strange data point may be nothing more than the luck of the draw. To further complicate matters, deviant observations may actually carry vital information, perhaps as "ideal" types of the units being studied. In short, as an alternative to fixing the data or discarding it, one needs statistical procedures that in some sense "accommodate" it (Barnett & Lewis, 1978).

More positively, one may decide on substantive grounds that the quadratic objective function is inappropriate. For example, suppose that one is regressing income on education. A least squares fit implies that individuals with unusually high or low incomes will have a disproportionate impact on the estimated regression parameters, especially if such individuals are also unusually high or low in education. Thus graduate students (presumably high in education and low in income) will have a far larger relative impact on the regression fit than, say, assembly line workers or secretaries. Note that more is involved than potential atypicality per se; the impact of any atypicality is magnified by the squaring process. Clearly, it would be useful to have estimation procedures that could provide alternatives to the quadratic objective function. M-estimators are one viable option.

M-ESTIMATORS OF LOCATION

Drawing on Li (1985, p. 291), the M-Estimator for the vector β is based on the objective function $\rho(t)$ and the data $(Y_1, X_1, \ldots, Y_n, X_n)$. It is the value of β, denoted $\hat{\beta}_m$ that minimizes

$$\sum_{i=1}^{n} \rho(y_i - x_i\hat{\beta}_m) = \sum_{i=1}^{n} \rho\left(y_i - \sum_{j=1}^{J} x_{ij}\hat{\beta}_j\right). \tag{7.1}$$

Equation 7.1 is a generalization of the conventional least squares objective function,[4] with $\rho(t)$ left unspecified. If the residuals are squared (i.e., $\rho(t) = t^2$), one has, as one M-estimator, ordinary least squares. If the absolute value is taken (i.e., $\rho(t) = |t|$), one has the least absolute residual estimator. These and other options specify the *particular* M-estimator being used, to which we will turn shortly.

It is often instructive to consider not just $\rho(t)$, but its derivative $\rho'(t) = \psi(t)$. If the goal is to minimize equation 7.1, taking the derivative of equation 7.1 with respect to $\hat{\beta}_m$, setting the result equal to zero, and solving leads to the desired result. Indeed, in the quadratic case, the intermediate result is the usual normal equations. Thus $\psi(t)$ figures centrally in the production of actual estimates, both in a generalization of the normal equations and as $\omega(t)$, a function of $\psi(t)$ and t, used as a weight in iteratively reweighted least squares. More will be said about estimation later.

In addition, the properties of M-estimators are often characterized in part through $\psi(t)$. For example, $\psi(t)$ figures in formal treatments of the *influence function* (e.g., Li, 1985, pp. 298-299), which will be addressed briefly below. In this introductory chapter, however, such uses of $\psi(t)$ are not essential and are discussed only in passing. Most of the central ideas can be addressed through the objective function and a few numerical illustrations.

Figure 7.1 plots the ordinary least squares (OLS) objective function against the values of residuals, called more generally deviation scores.[5] As the quadratic form implies, the weight given to deviation scores

increases at an increasing rate as the absolute value of the deviation scores increase.

Figure 7.2 shows the objective function when, instead of squaring the deviation scores, one takes their absolute values. The weights now increase linearly so that the larger deviation scores are given no special importance. While in Figure 7.1 the weights for deviation scores around 4 in absolute value were approaching weights of 6, the weights for those same scores in Figure 7.2 are a little over 3. Using the absolute value as the objective function leads to least absolute residual (LAR) regression and is another type of M-estimator.[6]

Figure 7.3 represents a compromise between least squares and least absolute residual regression. Up to the predetermined absolute value of a deviation score (e.g., |1.5|), the objective function is OLS. Beyond that value, the objective function is LAR. This is the Huber M-estimator.[7]

While LAR regression gives larger deviation scores less weight than ordinary least squares regression, there will be circumstances in which larger deviation scores will stem from suspect observations, and when, therefore, these observations need to be discounted. One discounting method can be seen in Figure 7.4. The bi-square is an M-estimator that weights deviation scores steeply up to some predetermined deviation score value (like the Huber M-estimator just described), at which point weights become constant. That is, beyond that predetermined deviation score value, larger scores are given the same weight as smaller scores.

Figure 7.5 shows a less "severe" discounting M-estimator, the Bell M-estimator. Like the bi-square, at some point increases in residual values do not translate into commensurate increases in weights, but full discounting to constant weights is only approached as a limit. Moreover, the shift to discounting occurs gradually.[8]

Statistical Performance Criteria

It should be clear from Figures 7.1 through 7.5 that M-estimators provide a rich menu of objective functions. But how can one choose between them? To begin, there is a set of statistical criteria that basically define how a "good" robust estimator should perform.[9]

First, it is clearly desirable for M-estimators to have the usual "large sample" properties of maximum likelihood estimators: consistency and asymptotic normality.[10] The OLS, LAR, and Huber estimators meet these criteria. The bi-square and Bell estimators will as well, as long as the distribution to which they are applied is strongly unimodal. If the

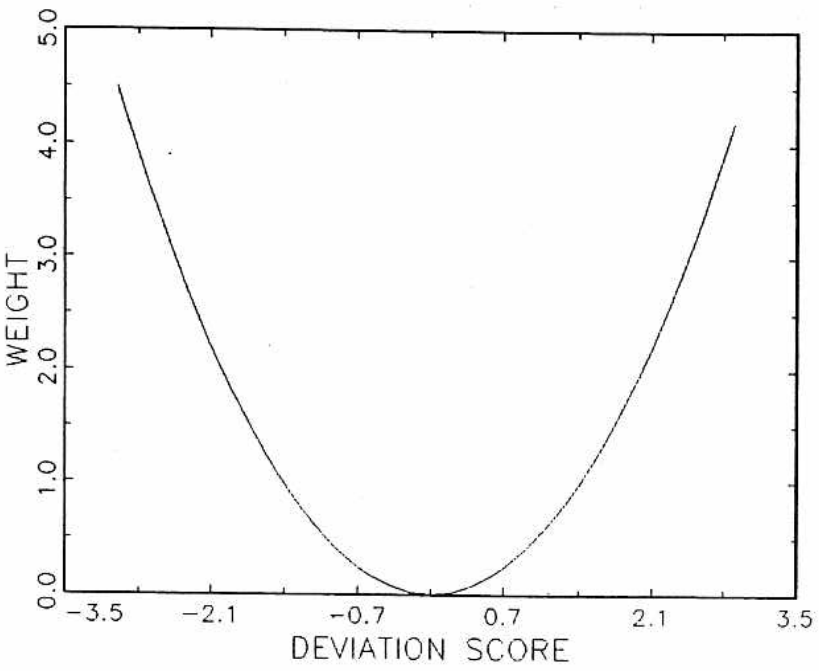

Figure 7.1 OLS objective function.

distribution is not strongly unimodal, the bi-square and Bell estimators may or may not be consistent and asymptotically normal, depending on technical considerations beyond the scope of this chapter. However, a good rule of thumb seems to be that the predetermined constant required by these estimators be kept relatively large (Wu, 1985). What this means is that the strong "discounting" of larger deviation scores does not begin until the larger deviations become quite large. Exactly what defines large is a tuning constant that may be manipulated (e.g., deviations larger than two in absolute value).

At the same time, however, it is very easy to make too much of good asymptotic properties. Small- to modest-sized data sets are common in the social sciences (e.g., $N < 200$), especially for large observational units such as organizations, cities, and countries. In samples of this size, it is typically very difficult to make any general statements about the

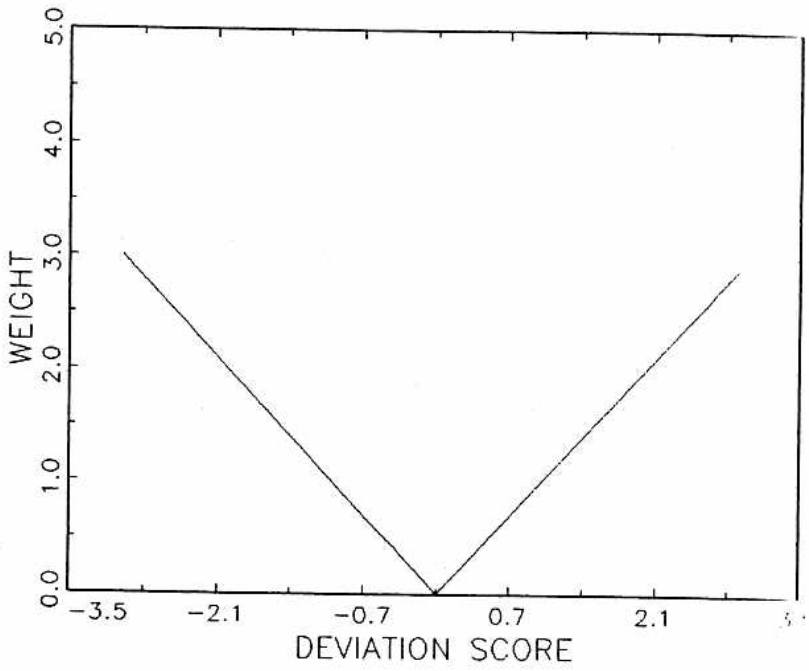

Figure 7.2 LAR objective function.

performance of M-estimators without specific assumptions about the distribution of the disturbance term. In short, asymptotic properties are often irrelevant.

Second, good M-estimators should be "resistant." Basically, a resistant estimator is relatively unaffected by a few rather deviant observations or many slightly deviant observations. Drawing heavily on Mosteller and Tukey (1977, pp. 350-352), consider the following 14 observations:

$$-6\ -5\ -4\ -3\ -2\ -1\ -.5\ .5\ 1\ 2\ 3\ 4\ 5\ 6\ .$$

Imagine another observation X that can be "moved" through the data, beginning with large negative numbers and ending with large positive numbers. For each increment in X, one calculates a summary measure

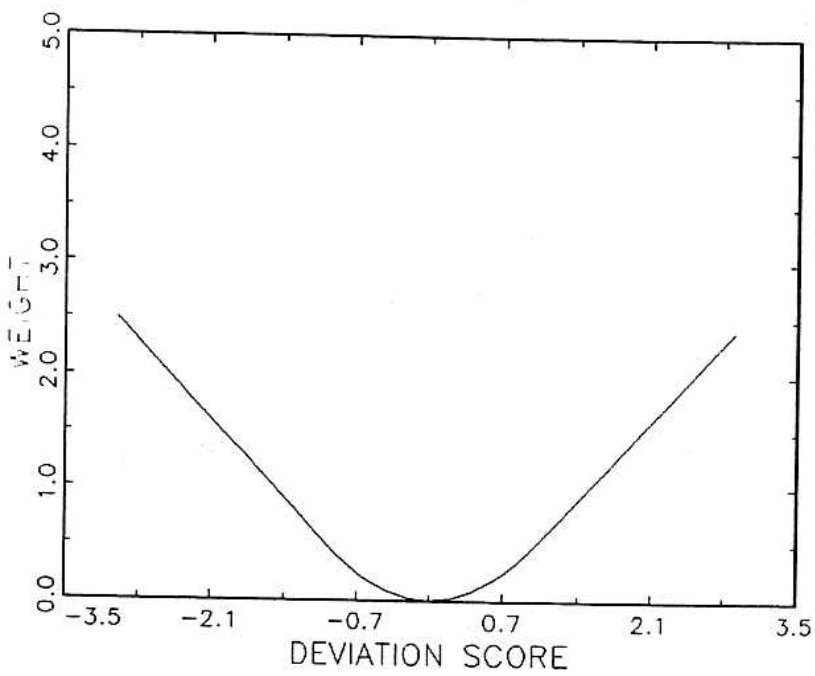

Figure 7.3 Huber objective function.

of location, so that it is possible to plot the measure against the changing value of X. If the summary measure is the mean, its value will increase in a linear fashion with increases in X. If the summary measure is the median, its value will not change until the increasing value of X exceeds $-.5$, will change linearly between $-.5$ and $.5$, and will not change for increased values of X larger than $.5$. In both cases, the plot of the summary measure against X conveys the degree of resistance. In both cases, one is, in effect, studying an influence curve. Figure 7.6 shows these influence curves for the mean and median.

Because the influence curve of the mean is unbounded, the mean is formally said to lack resistance. That is, the mean can be shifted any arbitrary amount with an arbitrarily large or arbitrarily small value of X. This implies that the mean is very vulnerable to anomalies in the data. The median is also formally said to lack resistance because of the sharp

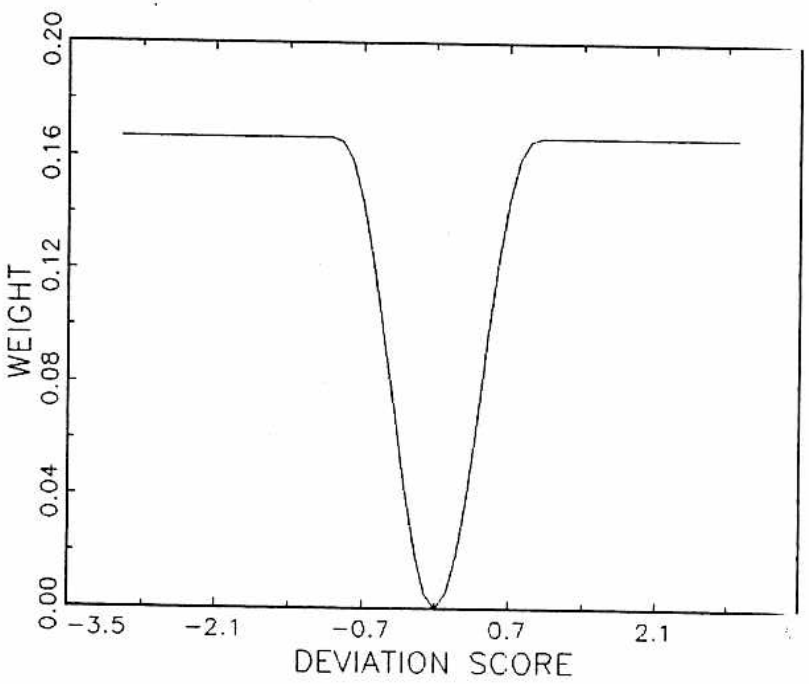

Figure 7.4 Bi-square objective function.

shift around the middle value(s) of the distribution. The point in both cases is that the summary statistic is rather easily bounced around when the data do not cooperate. However, because the shift in the median is bounded, the median is usually treated as if it were resistant.

A bit more formally, the influence curve actually shows how much the value of a particular estimator changes in response to infinitesimal changes in the underlying distribution. That is, one is able to examine how the estimate is altered by arbitrarily small changes in distribution. Mathematical statisticians find influence curves extremely useful, but for our purposes the overall message is that LAR, Huber, bi-square, and Bell M-estimators all considered resistant.[11]

Third, good M-estimators should have a high *breakdown point*. The idea of a breakdown point is closely related to the properties of influence curves. Suppose that some number of observations from a sample

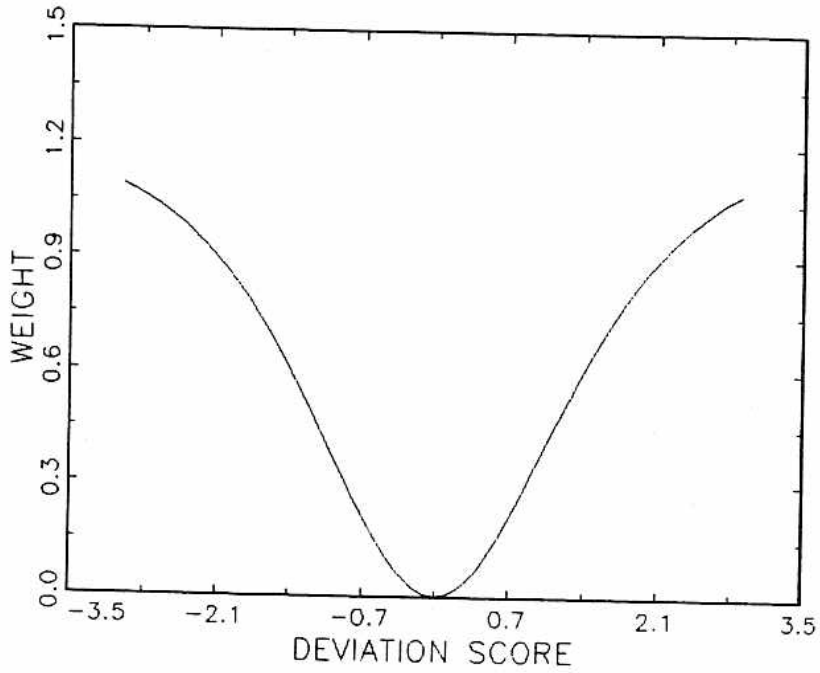

Figure 7.5 Bell objective function.

are arbitrarily replaced. Or suppose that to the given sample, some number of new observations is arbitrarily added. The breakdown point of an estimator is the smallest proportion of the sample that may be arbitrarily replaced or added, which may result in the estimate becoming unbounded (i.e., going off "to infinity"). For example, as Figure 7.6 suggests, the mean may be made arbitrarily large by adding a single, sufficiently large observation. In contrast, the median can be made to break down if half of the data is shifted. In practical terms, estimators with high breakdown points do not change dramatically in the face of large disparities between the assumed and actual distribution, including qualitative errors in the shape assumed. For M-estimators of *location*, all but the OLS estimator do quite well; they have relatively high breakdown points.

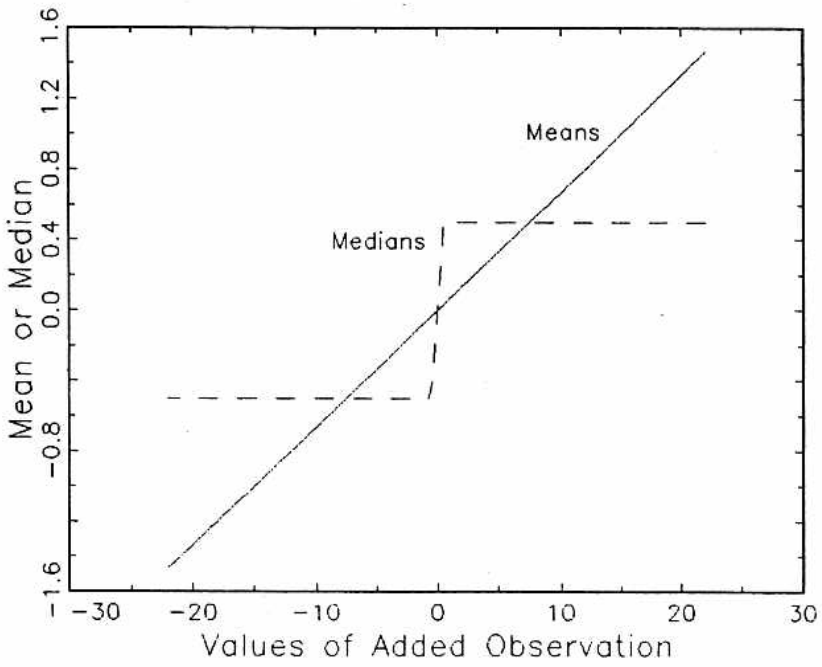

Figure 7.6 Influence curves.

The story is somewhat more complicated for *regression* M-estimators. The key idea is that each observation must be evaluated with respect to where it lies in the distribution of the response (dependent) variable *and* the joint distribution of the explanatory (independent) variables. Observations that are outliers on the response variable, and with respect to the joint distribution of the explanatory variables, are particularly problematic.

Consider, for example, the usual bivariate OLS regression. The regression line is fit in two-dimensional space, and outliers are distinguished by their location with respect to the bulk of the bivariate scatterplot. Regression M-estimators can discount the impact of outliers in the *y-direction*, but do not address outliers in the *x-direction*. In an important sense, only half of the problem is solved, even in principle. Thus regression M-estimators have low breakdown points (Rousseeuw

& Leroy, 1987, pp. 68-70). Estimators that discount outliers in both the y-direction and x-direction exist, but there are important trade-offs, which some feel make them problematic. There seems to be, for instance, a nearly inevitable trade-off between a high breakdown point and high efficiency. In any case, these alternatives to M-estimators are beyond the scope of this chapter.[12]

Fourth, good M-estimators are relatively efficient across a range of likely distributions. Recall that efficiency refers to the variance of an estimator's sampling distribution. Also recall that small (or *finite*) sample efficiency is determined by how an estimator performs in a sample of a particular (typically modest) size, while asymptotic efficiency is determined by how an estimator performs in a sample that becomes arbitrarily large. Under certain circumstances, optimal estimators may achieve the smallest standard error that is theoretically possible.

Unfortunately, evaluating the efficiency of M-estimators is complicated. To begin, optimal efficiency can only be achieved, even in theory, with respect to a specific distribution, and an estimator that may be optimal for one distribution may perform poorly for others (Wu, 1985, p. 350). As a fallback position, therefore, one typically focuses on relative efficiency, which is a standardized ratio of sampling distribution variances. Moreover, since it is rare for the data's distribution to be known, it makes sense to pick an M-estimator whose relative efficiency (compared to other M-estimators) is high across a range of possible distributions. In the end, however, the only general conclusion seems to be that the least squares estimator performs worst. That is, even with modest deviations from the normal distribution, relative efficiency falls off dramatically. Among the other M-estimators, overall conclusions depend on the particular set of distributions being considered.[13]

Finally, good M-estimators should be practical: relatively easy to compute, useful for a variety of data problems, and comprehensible to the mere mortals who will have to use them. All M-estimators for robust regression are reasonably practical.

To summarize, a review of statistical criteria for robust regression is primarily an exercise in OLS bashing. Selection among the remaining M-estimators seems too often to be data specific and dependent on judgment calls about the relative importance of different performance characteristics. In particular, there is often a trade-off between the breakdown point and efficiency. Moreover, much of what we know about the performance of M-estimators depends on asymptotic proper-

ties, which may be misleading for samples of the size often available to social scientists. An important implication, to which I shall return, is that one may well benefit from trying a variety of approaches and proceeding in a more inductive manner than is commonly recommended (see, for example, Leamer, 1983).

Substantive Criteria

While it may be difficult to provide general guidance on the choice of an M-estimator from statistical performance criteria alone, there will often be times when choices between different M-estimators can be made on substantive grounds. Recall the earlier example in which income was regressed on education, and the relative importance of data on graduate students was considered. Depending on how one chooses to weight the deviation scores, the regression line estimated can differ substantially. If larger deviation scores are discounted relative to smaller deviation scores, the regression line will more closely summarize the experience of more typical individuals. That is, information from more typical individuals is treated as more important than information from less typical individuals.

If one has reason to suspect that atypicality on the average results from some anomaly, the discounting may make sense. Alternatively, there may be no reason to differentially weight atypical observations; indeed, there may be circumstances when they should be given extra weight. For example, perhaps observations near the center of the scatterplot represent cases in which the available measures failed to record more extreme values. Thus low income individuals may underreport their income for fear of losing eligibility for various kinds of transfer payments; or smaller municipalities with less professional public servants may routinely fail to record incidents, such as reported crimes, that are later aggregated as the official statistics for the locale. The point is that there will be situations in which, on substantive grounds, typical observations may be less credible than atypical ones. Then, the typical observations should be downweighted, or at the very least, not given extra weight. In short, before proceeding it is vital to consider objective functions such as those shown in Figures 7.1-7.5 and decide which makes the most substantive sense.[14]

Computation

Regression M-estimators are perhaps most easily computed with any software that contains procedures for iteratively reweighted least squares. SAS, BMDP, GAUSS, and PC-ISP are examples.[15] One begins with conventional OLS estimates and then weights the data (the response variable, the explanatory variables, and the vector of 1's for the intercept) by a function of residuals. The particular function used depends on the M-estimator being employed. OLS is then applied to the weighted data. Again, residuals are calculated, new weights are constructed, and the data are reweighted. OLS is then applied a third time. The OLS estimation and the reweighting is continued until the estimates converge. Table 7.1, reproduced from Li (1985, p. 293), shows for popular M-estimators the objective function $\rho(t)$, the derivative of the objective function $\psi(t)$, and the weighting function $\omega(t)$.

With the exception of OLS and LAR regression, a scale parameter (much like σ^2 in OLS regression) also needs to be estimated *along with* the usual regression parameters. That is, a scale parameter is required as part of the iteration process; all of the residuals are divided by (i.e., scaled by) the scale parameter. This presents no special difficulties, although there is some debate about what scale parameter estimator should be used. Details can be found in Li (1985, pp. 300-310).

Unfortunately, the issues are far more complicated when one turns to statistical inference. First, just as in the usual formulas for OLS regression, a scale parameter is required for calculation of the standard errors. However, a key motivation for robust regression is concern about outliers, and that same motivation applies to estimates of scale. Hence, one needs a sensible robust scale estimator, and there are many possible candidates. For example, a linear objective function leads to the mean absolute deviation (MAD) scale estimator. Yet there seems to be no consensus about which is best, and the difficulties caused by a number of unresolved technical matters. Second, statistical inference requires that the sampling distribution of the estimates be known. For large samples one can rely on asymptotic normality, but for small samples the sampling distribution is almost certainly not known and may well be a very long way from normal. These and other difficulties make statistical inference for M-estimators problematic (Li, 1985, pp. 300-301; Wu, 1985, pp. 365-363, 367).

Perhaps the best approach, therefore, relies on resampling methods such as the bootstrap (see, in particular, Chapter 8, this volume). The

Table 7.1 M-Estimators for Regression

Estimator	$\rho(t)$	$\psi(t)$	$\omega(t)$	Range of t
OLS	$\frac{1}{2}t^2$	t	1	$\lvert t\rvert < \infty$
LAR	$\lvert t\rvert$	$sgn(t)$	$\frac{sgn(t)}{t}$	$\lvert t\rvert < \infty$
Huber[a]	$\frac{1}{2}t^2$	t	1	$\lvert t\rvert \le k$
	$k\lvert t\rvert - 1/2k^2$	$k\,sgn(t)$	$k/\lvert t\rvert$	$\lvert t\rvert > k$
Andrews[a]	$A^2[1-\cos(t/A)]$	$A\sin(t/A)$	$A/t\sin(t/A)$	$\lvert t\rvert \le \pi A$
	$2A^2$	0	0	$\lvert t\rvert > \pi A$
Biweight[a] (bi-square)	$\frac{B^2}{6}\{1-[1-(t/B)^2]^2\}$	$t[1-(t/B)^2]^2$	$[1-(t/B)^2]^2$	$\lvert t\rvert \le B$
	$\frac{B^2}{6}$	0	0	$\lvert t\rvert > B$

SOURCE: David C. Hoaglin, Frederick Mosteller, and John Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes*. Copyright © 1985. John Wiley & Sons. Reprinted by permission of John Wiley & Sons, Inc.

NOTES: a. The illustrative example ρ-functions and ψ-functions use $k = 1$ for the Huber, $A = 1/\pi$ for the Andrews and $B = 1$ for the biweight.

basic idea of the bootstrap is to treat the data set as a population. Certainly, if the data were sampled properly by probability procedures, the data set will well represent the population, within sampling error. Then, one takes bootstrap samples from the data set by selecting single cases at random *with replacement*. (Each bootstrap sample is the same size as the data set.) From bootstrap sample to bootstrap sample, parameter estimates will vary. In effect, the sampling distribution of the estimator is being empirically generated. Statistical inference then follows naturally. In the application to follow in the fifth section, bootstrapping is employed.

AN ILLUSTRATION

Before launching into a *real life* application with all of its complexities and uncertainties, a far more simple illustration may perhaps prove useful. In an article on Adolphe Quetelet, Stone (1988) included data on the number of births and deaths by time of day for a particular

Table 7.2 Births and Deaths in Brussels by the Hour

Hour	*Births*	*Deaths*	*Hour*	*Births*	*Deaths*
1	142	228	13	94	257
2	173	253	14	97	233
3	130	230	15	88	217
4	122	242	16	91	237
5	120	231	17	104	281
6	111	213	18	100	233
7	112	217	19	121	204
8	99	248	20	97	194
9	88	207	21	133	199
10	130	228	22	115	220
11	137	311	23	224	243
12	48	110	24	4	14

hospital in Brussels. The number of births by hour covers a 30-year period in the nineteenth century, whereas the number of deaths dates from 1811 to 1822. Table 7.2 shows these data.

Figure 7.7 (constructed with the statistical package STATA) shows the scatterplot for deaths (y) and births (x), along with the univariate boxplot for each. Twenty-two of the observations are clustered and show little association. Two observations (for noon and midnight) are dramatically smaller in both the y-direction and x-direction. With these two included, there is obviously a positive association in the data. It is difficult to know what to make of the two apparent outliers without being a lot more familiar with how the data were recorded and collected. However, since there is no apparent biological reason for the outliers, hospital practice or data collection are implicated.

Table 7.3 shows four regression estimates for the bivariate relationship shown in Figure 7.7. The first is the ordinary least squares estimate. Roughly speaking, there is a one-to-one increase of deaths with births. The intercept is approximately 100. Both coefficients are statistically significant at the (two-tailed) .05 level for a null hypothesis of zero. However, given the small sample and real questions about the disturbance distribution, both tests are probably not very useful. Next (moving to the right) are shown the results for least absolute residual regression. The estimated relationship is, in effect, rotated clockwise: The slope is cut by about 50% and the intercept is increased by about 50%. No standard errors are presented because there is no really convenient way of getting them.[16] Next are shown the Huber regression

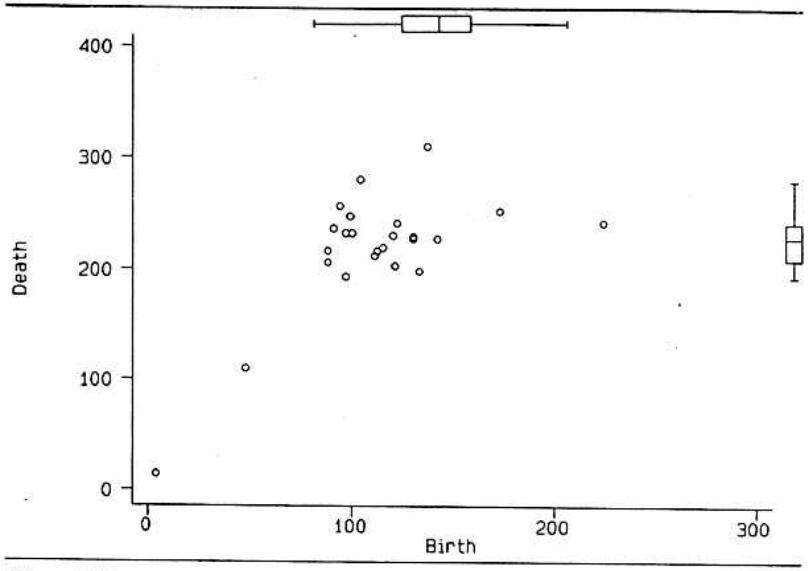

Figure 7.7

results and little changes. However, standard errors are provided, and both coefficients are *statistically significant*, if one assumes that the disturbance distribution is normal. In short, both robust regression M-estimators tell a rather different story compared to ordinary least squares.

Recall, however, that regression M-estimators only address outliers in the y-direction. The final column in Table 7.3 shows the results for least median squares regression (calculated with PROGRESS by Rousseeuw & Leroy). Least median squares regression (Rousseeuw & Leroy, 1987) fits the regression line (hyperplane) by minimizing the median of the squared residuals rather than the sum of the squared residuals; that is, the criterion is a robust measure of location rather than a very nonrobust sum. While the details are beyond the scope of this chapter, outliers in both the y-direction and x-direction are taken into account. Note that the estimated relationship is now rotated further in a clockwise direction. Compared to the robust M-estimators, the slope is decreased by about half, and the intercept is increased by about a third. Indeed, there now appear to be virtually no relationships between births and deaths. No standard errors are available. Nevertheless, the

Table 7.3 Quetelet Birth-Death Data: Results for Three Regression M-Estimators and LMS Regression

	OLS	*LAR*	*HUBER*	*LMS*
CONSTANT	114.9	159.4	169.9	195.8
	(26.31)	*	(18.5)	*
COEFFICIENT	.92	.54	.47	.26
	(.22)	*	(.16)	*

NOTE: *Standard error not computed.

vulnerability of regression M-estimators to outliers in the x-direction is apparent.

In summary, the illustration makes clear that regression M-estimators can make a difference. However, their small sample properties are typically unknown in just those instances when they are most likely to be needed. In addition, outliers in the x-direction are ignored. We turn now to a far more realistic application.

AN APPLICATION

While robust regression represents a particular set of estimation procedures, equally important is the underlying data analysis perspective. At each step in the process, from research design to reporting results, one proceeds as if Murphy's Law applies. This means that as many assumptions as possible are made problematic, and, where possible, efforts are made to protect the analysis. This also means being explicit and conservative about what may be learned. As an illustration, I present below an evaluation of an effort in Alameda County, California, to more effectively prosecute narcotics cases (Greenspan, Berk, Feeley, & Skolnick, 1988).

The Program

On January 10, 1985, Oakland's Assemblyman Elihu Harris introduced legislation in the State Assembly that was intended to coordinate and enhance law enforcement efforts to control drug use in Alameda County. Particular attention was directed toward the courts. The bill assumed that more effective and efficient prosecution of narcotics cases could lead to a reduction in drug crimes and drug-related crimes. A

number of interventions were proposed, including an oversight "Targeted Urban Crime Narcotics Task Force" and additional financial support for the county's courts, Prosecutor's office, Public Defender's office, probation department, and crime laboratory. The bill was approved in July of 1985, and program funding became available on October 1, 1985.

For the present purpose, I will focus on whether the legislation made the sanctioning process more effective, and on a particular outcome measure: the number of offenders incarcerated. I will not address the ultimate impact of the program on crime or other kinds of outcomes such as efficiency (e.g., how fast cases were processed). Readers interested in the substantive issues should consult the evaluation completed by Greenspan and her colleagues (1988).

Under the Alameda Program, there were essentially two routes by which drug offenders could be incarcerated. Offenders could be sentenced by the court after a conviction (or pleading) or be sent to prison or jail for having violated probation. For the first route, the state legislation supported the use of a *team approach* to prosecution, in which all drug-related offenses were handled by a specialized group of Deputy District Attorneys, under the direction of a *coordinator*. The team was given sufficient staff to try at least two cases simultaneously.

For the second route, an effort was made to orchestrate better the probation revocation process so that drug offenders who violated the conditions of their probation would be swiftly incarcerated. Offenders found violating a condition of their probation were required either to serve the original sentence imposed or a new sentence if the original sentence had been suspended. One key advantage of incarceration through probation revocation was that the standards of proof are lower than in court trials. This meant that it was often more effective to simply "violate" an offender than to go through the trouble of trying the offender for a new crime. A single member of the drug prosecution team was made responsible for revocation process.

Despite the face validity of the program, it was not at all clear that it would work. For example, the greater use of probation revocation might divert "good" drug cases away from the usual channels. Thus, while revocations might increase, convictions might decline. Alternatively, the District Attorney's office might have to use the additional resources to aggressively pursue a small number of difficult cases, with the bulk of the caseload unaffected. Anticipating such questions, the legislation required a program evaluation.

The Research Design

The legislative requirement for an evaluation was not matched by a great deal of insight into what a sound evaluation would entail. For a variety of reasons, therefore, the strongest design that could be implemented was an interrupted time series from official statistics. In brief, data on cases processed by the county's court were available from a management information system (called CORPUS) used to monitor the processing of cases forwarded to the District Attorney's office. These data were examined and from tapes provided by the county, a longitudinal file was constructed organizing key outcome variables by quarter. There were 12 quarters of data, with the intervention falling in the eighth quarter. I will focus on the number of drug cases sanctioned over those 12 quarters.

Figure 7.8 shows with a broken line and open squares the time series for the number of drug cases in which the offender was sanctioned (i.e., sent to prison or jail). The number of drug offenders sanctioned by quarter ranges from about 70 to about 350, but there is clearly an upward trend, beginning before the eighth quarter. Moreover, since the steps from arrest to sentence may take months, cases sanctioned in the eighth quarter were largely processed before the program began; that is, program effects on sanctioning should appear after the eighth quarter. And in this case, the trend begins in about the fourth quarter, especially if the downward spike in the eighth is considered aberrant.

Note how one could have been misled by a pretest/posttest design. A comparison between the number of cases sanctioned in the seventh quarter, for example, and the tenth quarter would have revealed a dramatic gain of about 100 cases, but would have neglected the positive trend beginning well before the program was launched. This illustrates simply an important principle in a robust approach: One can reduce dramatically the difficulties faced during data analysis by anticipating possible difficulties in one's research design.

In this instance, the design can be further strengthened by adding a comparison group not subject to the program. Figure 7.8, therefore, also shows a time series for the number of theft cases sanctioned (in which no drug offenses were involved). As with the drug cases, there is a general upward trend, although it seems to begin a bit later.[17] This supports the speculation that there may be no distinct program impact because the increase that both series share must be driven by common or correlated causes.

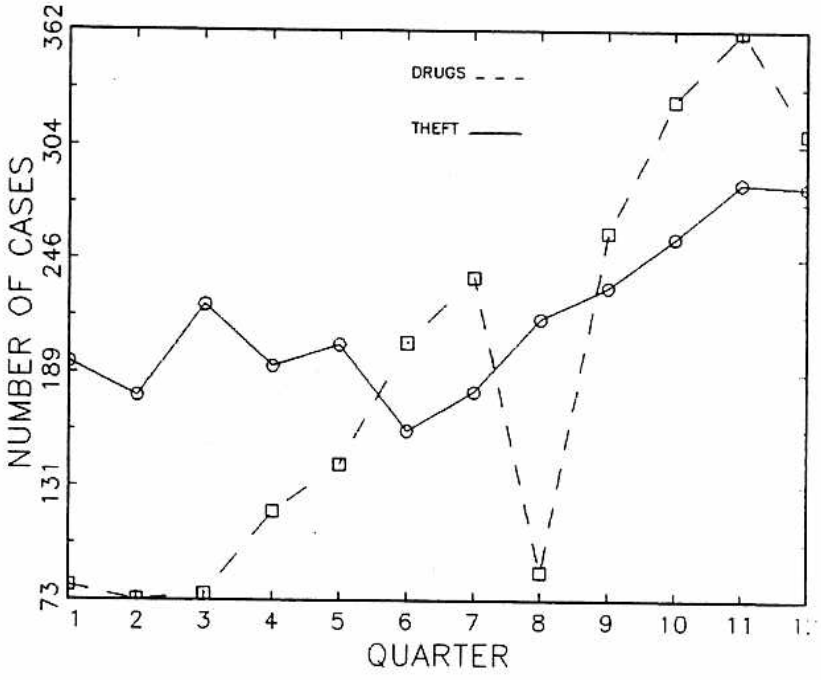

Figure 7.8 Cases sanctioned.

The search for common or correlated causes leads directly to a concern with the number of drug and theft cases entering the system; more cases coming in the front door must generate more cases going out the back door. This leads to standardizing the number of sanctioned by the number of arrests. That is, one may control for the number of cases entering the system by simply calculating the proportion of arrests for which sanctions were applied.

Figure 7.9 plots, therefore, the proportion of final dispositions in which sanctions were imposed. The proportions range from a low of about .11 to a high of about .40. Thus, for example, a high of about 40% of the theft dispositions involved incarceration. Perhaps the major conclusion from Figure 7.9 is that the drug series seems to be gaining on the theft series.

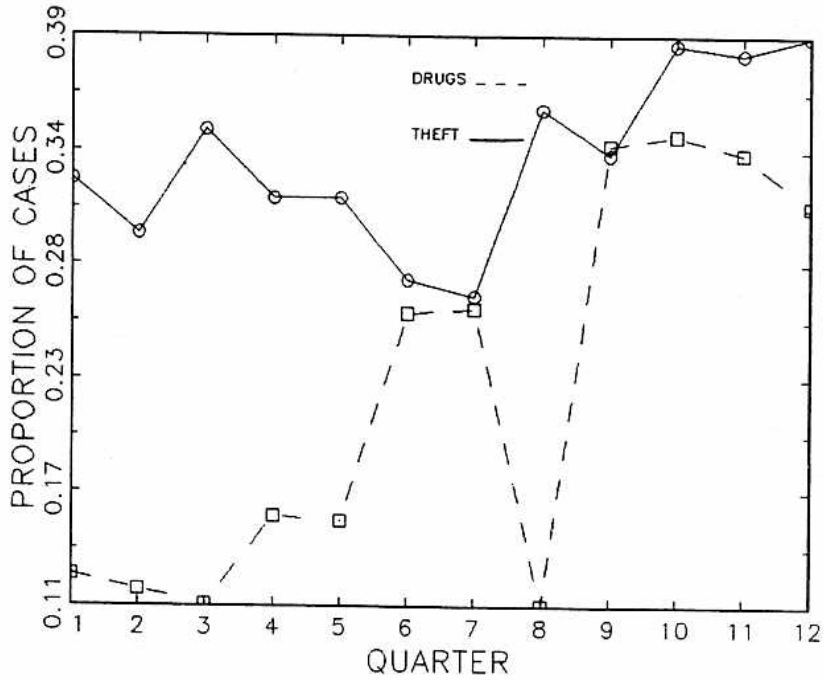

Figure 7.9 Cases sanctioned.

Unfortunately, it is still difficult to disentangle the long-term relative gains for the drug series, beginning in the sixth quarter, from gains after the eighth quarter that may be attributable to the program. Inferences are made especially difficult because of the possible outlier in the eighth quarter for the drug time series. Alternatively, the eighth quarter may be reasonably atypical of quarter-to-quarter variation, and the proportions for the sixth and seventh quarters may be atypical. In short, there seems to be unusual variation in the drug time series in the middle of the observational period, but its causes and consequences are unclear. In any case, the importance of visual displays should be apparent; there is no substitute for a careful examination of one's data before statistical analysis is begun.

Statistical Analysis

As a way of capitalizing on the information produced by the research design, an outcome variable was defined as the difference between the proportions of drug and theft cases sanctioned. Using proportions standardized for the number of cases overall reaching final disposition controls for the number of cases entering the system, whereas differencing the two time series controls for common or correlated causes.[18] From the perspective of robust data analysis, differencing has the asset of requiring that no parameters be estimated.

As for any time series, there is also reason to be at least suspicious about autocorrelation within the differenced time series. This suggests the need for some kind of autoregressive formulation. Were there more than a suspicion, one might choose to longitudinally difference the (already cross-sectionally differenced) series. However, since "over-differencing" can lead to biases in the analysis of time series data, it is probably more sensible to allow the amount of difference to be an empirical question.

Figure 7.10 shows with a broken line the difference between the standardized drug and theft time series. Using that differenced series as the response (dependent) variable, an OLS model was fitted using one-period lagged values of the response variable and a dummy variable for the treatment, coded 0 through the eighth quarter and 1 thereafter. The lagged response variable was introduced to control for any first order serial correlation, and the dummy variable was introduced to estimate any treatment effects. The solid line in Figure 7.10 represents the *predicted values* from the OLS regression.

Figure 7.10 suggests that there was a treatment effect, and, from the upper panel in Table 7.4, we see that although the autoregressive component is not important, the estimated treatment impact is. The drug series gains 9% on the theft series after the program is introduced (beginning in the ninth quarter). Looking back at Figure 7.9, one can see that a 9% increment is nontrivial, given base sanctioning rates between 11% and 40%. Put another way, about 20% of the pretest final dispositions in drug cases include incarceration. The OLS estimates of the program's effect suggest an increase in that figure to about 30%, or about a 50% relative improvement. However, under the t distribution (given the small sample size), the effect is not statistically significant at the .05 level for either a one-tailed or a two-tailed test.

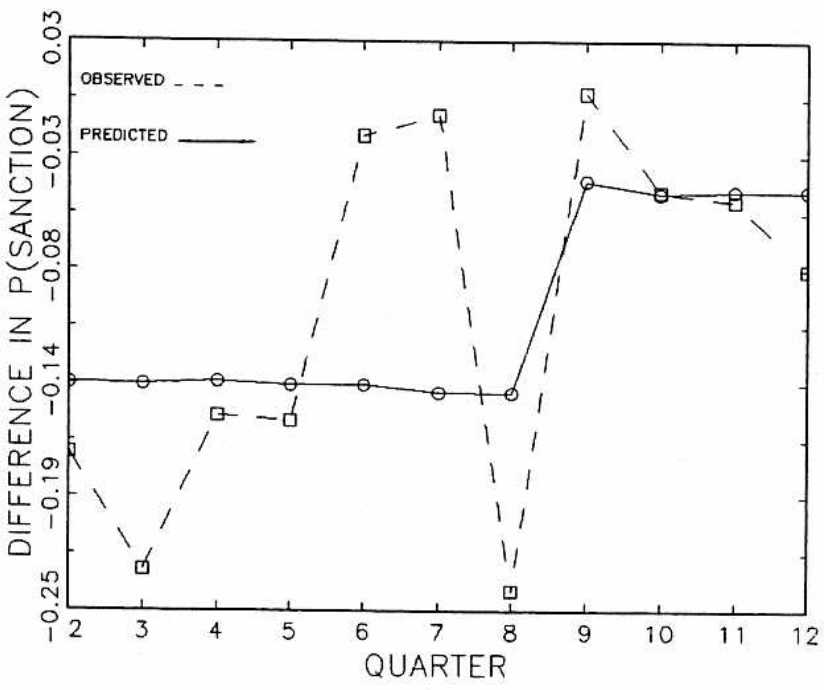

Figure 7.10 Cases sanctioned.

There are, however, ample grounds for being uneasy with these results. First, because of the small sample, conventional significance tests require that the disturbance term have a normal distribution. This is formally impossible since the response variable is the difference between two proportions. Indeed, there are not even grounds for assuming that the disturbance term has a symmetric distribution, in part because the drug series has several observations close to zero.

Second, there is no reason to assume that the events making up each proportion are independent. Indeed, given the bureaucratic environment in which the cases were processed, the events are probably clustered by *spells*. That is, there is a spell of great concern about crowded court dockets, followed by a return to business as usual. During a spell of concern, cases are processed faster than at other times, implying serial correlation between the length of time between events and,

Table 7.4 Regression Analysis for the Impact of the Program

	Ordinary Least Squares Regression		
Variable	Coefficient	Standard Error	*T* Value
Intercept	−0.14	0.05	−2.82
Program impact	0.09	0.05	1.79
Lagged outcome	−0.03	0.29	−0.09
	Bootstrapped Least Absolute Residual Regression		
Variable	Coefficient	Standard Error	*T* Value
Intercept	−0.21	0.05	−4.33
Program impact	0.16	0.06	2.68
Lagged outcome	−0.30	0.27	−1.10

therefore, the events themselves. While the lagged response variable may well "soak up" any correlations between the proportions over time, any correlation among the events making up the proportions remains a potential problem, which undermines conventional statistical inference.

Third, the very large residuals for quarters six, seven, and eight are grounds for concern. Perhaps the results are being inappropriately dominated by these three quarters. Within the pretest period at least, the data for quarters six and seven look particularly aberrant. In short, there is good reason to worry about the quadratic objective function and a rationale for trying a robust alternative.

A priori, there seems to be no reason for significantly downweighting the larger residuals relative to the smaller ones. That is, there is no reason to suspect that the three largest residuals result from a measurement or design error. This suggests ruling out any of the redescending M-estimators in favor of least absolute residual regression. Recall that the objective function for least absolute residual regression weights the residuals in a linear fashion.

Recall that LAR regression can be easily undertaken with iteratively reweighted least squares (Li, 1985, pp. 305-310). Basically, before each least squares "pass," all of the data for each case (including the column of 1's associated with the intercept) are multiplied by the square root of the inverse of the absolute value of the residual for that case.[19] Unfortunately, it is not clear how best to calculate the standard errors directly (Li, 1985, pp. 300-301), and for these data, the small sample precludes any reliance on a normal asymptotic distribution for the parameter estimates (Amemiya, 1985, p. 75).

In response, the entire procedure was bootstrapped using resampling techniques appropriate for autoregressive time series models (Efron & Tibshirani, 1986, p. 65), but applied to M-estimators (Efron, 1982, pp. 35-36). Bootstrapping has a number of strengths, including the ability to represent all sources of instability, not just those addressed by conventional significance tests. However, it is necessary to assume that one's sample is truly representative of some theoretical population because, in a very real sense, the sample is being treated as a population. In the case of historical data such as these, the best that one can typically do is assert that if the underlying historical process in principle produces a population of realizations with the same properties as those observed in the given sample, the bootstrapped sampling distribution is appropriate. Although this may seem like a long stretch, the same argument basically applies to conventional statistical inference used on historical data. For both, the population is a hypothetical set of realizations from a given historical process.

Table 7.4 reports in the lower panel the LAR regression results based on 1,000 bootstrapped samples (estimated using GAUSS). Note that while the standard errors are basically unchanged, the treatment effect has approximately doubled. This doubling translates into a nearly 100% increase in the proportion of cases sanctioned. Note also that the *t* value is now well over 2.00. However, while the autoregressive coefficient has increased substantially, it is still not much larger than its standard error.

Taking the point values of all of the regression coefficients seriously for the moment, Figure 7.11 shows the goodness of fit. Clearly, the impact of the residuals for sixth and seventh quarters has been significantly reduced under the linear objective function. The result is lower estimates of the pretest predicted values leading to a larger estimated increment during the posttest period.

Figure 7.12 shows the bootstrapped sampling distribution for the treatment coefficient, based on 1,000 bootstrap samples. The distribution shows some skewing to the right, which, as noted earlier, is not surprising given the distributions of the theft and drug proportions; the right tail and left shoulder are heavy. Yet, because the distribution falls off more quickly on the high end, the mean, mode, and median are about the same (about .16). Ninety-five percent of the estimates fall between .03 and .33, which defines the 95% confidence interval, and 99.2% of

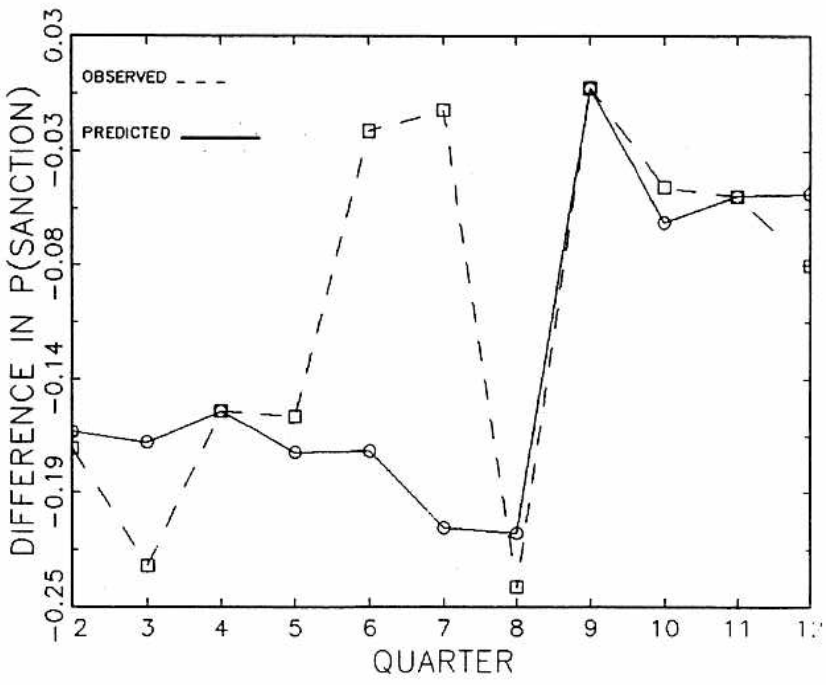

Figure 7.11 Cases sanctioned.

the estimates are above 0. Clearly, it is very unlikely that there is no posttest increment.

The story could have been different under conventional inference. If the t distribution had been applied (given the small sample size), the 95% confidence interval would have been between –.02 and .30. Under a two-tailed test, therefore, one would have failed to reject at the .05 level the null hypothesis of no treatment effect. However, the treatment effect would have been statistically significant at the .05 level for a one-tailed test.[20]

To summarize, there was ample reason to be suspicious about the OLS estimates. Because of the quadratic objective function, large residuals immediately before the intervention could have been distorting the results. LAR regression, coupled with bootstrapping, led to more plausible estimates of the treatment effect and to the conclusion that

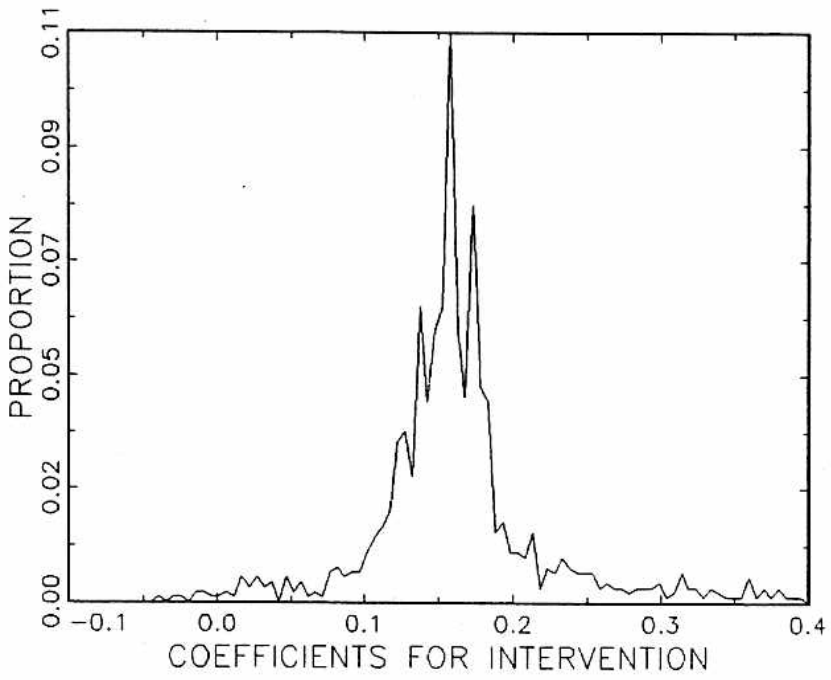

Figure 7.12 Distribution of estimates.

sanctioning increased after the program was initiated. At the same time, attribution of cause is always fraught with risk, and it was only after considering a range of other possible explanations for the posttest increase that policy recommendations were made (Greenspan et al., 1988).

CONCLUSIONS

Choosing a proper regression estimator requires that a pair of complementary judgments be made: What objective function makes substantive sense, and what estimator(s) has the best statistical properties? When the quadratic objective function makes substantive sense and when the normal distribution is closely approximated by the observed

disturbance distribution, one has, regardless of sample size, the best of all possible worlds: convenient least squares estimators with excellent statistical properties that match the questions being asked. With large samples, one is able, as a fallback, to capitalize on a number of handy asymptotic properties, even if the observed disturbance distribution is a long way from normal.

However, estimators relying on the quadratic form are fragile, and one may be seriously misled when other objective functions are more appropriate. Moreover, in the presence of outliers, least squares regression may lead one astray. Finally, least squares estimators may be well short of optimal, especially in small samples, if the observed disturbance distribution is not normal.

Given the widespread availability of cheap computing and the relative ease of computing M-estimators, M-estimators should be applied, at least as a supplement to OLS, whenever there is any doubt about OLS. When M-estimators and OLS (which, technically, is also an M-estimator) produce the same substantive story, all may be well. When they differ, the data must be carefully examined for these reasons. It should then be possible to make an informed decision about which set of results is most plausible.

More generally, there is an enormous gap between what sociologists know and what sociologists need to know to use properly the rich set of tools statisticians have provided. To meet the assumptions required for instructive statistical analyses, sociologists must have a reasonably accurate understanding of the substantive phenomena in question and have access to data sets unsullied by significant measurement and design errors. In other words, the data need to be an accurate reflection of some underlying social process whose general properties are known. Stated in these bald terms, it should be apparent that, in principle, almost all quantitative research in sociology is suspect and that statistical analyses of sociological data should be designed from a robust perspective. In practice, this means being very cautious about what may be learned from a given data set and regularly applying techniques that minimize the risk of being seriously misled. Routine application of a robust perspective could dramatically improve quantitative research in sociology.

NOTES

1. There are, however, important differences of opinion about the relative merits of robust estimators, and as further developments occur, the comparative advantages of different estimators may change. For example, Koenker and Portnoy (1987) have recently proposed a very interesting L-estimator for linear regression that asymptotically is the same as the Huber M-estimator (see below), but, unlike the Huber M-estimator, is scale invariant. For an interesting set of exchanges on such issues, see Draper (1987) and the comments that follow.

2. Fortunately, excellent discussions can be easily found elsewhere (e.g., Hampel et al., 1986; Li, 1986; Wu, 1985). Especially interesting are "generalized" M-estimators (Hampel et al., 1986, chap. 6; Li, 1985; Welsch, 1980) and certain S-estimators (Rousseeuw & Leroy, 1987) that address the impact of deviant values among one's *explanatory* variables as well as the impact of deviant residuals. I will have more to say about this issue below.

3. Since the "objective" of least squares procedures is to minimize the sum of the squared residuals, the function that does this is called an *objective function*.

4. It is also easy to show that M-estimators are a *slight* generalization of conventional maximum likelihood estimators (Hampel, 1986, p. 36). The M in M-estimator refers to its maximum likelihood roots.

5. These were drawn from a rectangular distribution, but any set of values covering a reasonable range would suffice. The distribution of the input to the objective function is irrelevant to the shape of its output.

6. The mean is the location measure that minimizes the sum of the squared deviation scores. The median is the location measure that minimizes the sum of the absolute values of the deviation scores. In effect, therefore, ordinary least squares regression fits a set of conditional means to the data while least absolute value regression fits a set of conditional medians. Consequently, many of the comparative merits of means and medians carry over to the two regression generalizations.

7. Where one sets the cutoff point is basically a judgment call, although there are diagnostics that may help.

8. Both the bi-square and Bell estimators belong to a class of "redescending" estimators because the derivative of the objective function, $\psi(t)$, first increases and then decreases to zero. There are also location estimators that give no weight whatsoever to deviation scores beyond a certain size by literally dropping them from the analysis. However, these are not M-estimators. The "trimmed mean" is one example (Rosenberger & Gasko, 1983, pp. 307-312).

9. The discussion that follows on desirable properties of M-estimators borrows heavily from Wu's excellent exposition (1985, pp. 325-327, 344-356).

10. As with all maximum likelihood estimators, one must assume that for the observed distribution in question (e.g., for the particular response variable), each observation behaves as if it were randomly and independently sampled from a particular distribution. However, this is not as restrictive as it might seem because the independence is *conditional* upon the values of the distribution's parameters and whatever conditioning variables are being used. In the case of linear regression, for example, the independence is found in the disturbance term (not the dependent variable per se), which represents the *conditional* distribution of the dependent variable around the regression hyperplane.

11. Technically, resistance requires that the derivative of the objective function ($\psi(t)$) be continuous and bounded. While the derivative of the objective function for the LAR estimator is discontinuous (at zero), the discontinuity is unimportant in practice.

12. One alternative, *bounded influence regression*, is briefly described in Li (1985, pp. 324-328). A major drawback is that the breakdown point is a decreasing function of the number of explanatory variables; in the very instances when the opportunities for outliers are great, the breakdown point is relatively low (Rousseeuw & Leroy, 1987, p. 13). Another alternative is "least median of squares" regression, which minimizes the median of the squared residuals instead of the sum. While least median of squares regression has a high breakdown point (indeed, it achieves the theoretical maximum of 50%, it gives up some efficiency (Rousseeuw & Leroy, 1987, sect. 4). Still, one must keep in mind that efficiency is calculated with respect to a particular distribution, typically the normal, and the relative efficiencies of estimators can easily change if the assumed distribution is incorrect. An unusually clear exposition of these issues can be found in Rousseeuw and Leroy (1987), and software for their preferred estimators is easily obtained from the authors.

13. Readers with good statistical backgrounds and a particular interest in efficiency might well want to work through Chapter 6 in Hampel et al. (1986). However, it is not clear to me that much of genuine practical significance will be learned.

14. This is especially important in applied work. See, for example, Berk and Cooley (1987) and Berk (1988).

15. Some software have routines designed especially for certain M-estimators. For example, SAS has a procedure for LAR regression, and PC-ISP has procedures for LAR and Huber regression estimators. PROGRESS does least median squares regression and trimmed estimator via weighted least squares.

16. This is because the second derivative of the objective function is not defined. Put another way, the underlying disturbance distribution (the Laplace distribution) is not continuous. Bootstrap methods are employed below.

17. Theft cases were chosen because they are common and in some ways similar to drug cases. But the basic point is that the program was directed at drug cases and not theft cases. Time trends that both share, therefore, cannot be attributed directly to the program.

18. The differencing is identical to inserting into a regression analysis a dummy variable for every time period but one, which is common in analyses of pooled cross-sectional and time series data within an analysis of covariance perspective (Hsiao, 1986, pp. 29-32). It is also closely related to the notion of cointegration for time series data (Granger & Newbold, 1986, pp. 224-226). Note that differencing does not assume a constant disparity between the two series. Shared effects that vary over time are removed.

19. There is a tendency for iteratively reweighted least squares, when applied to LAR regression, to produce one estimated residual very close to zero. Should this cause the software to abort before convergence is reached, a vary small number (e.g., .00001) can be added to each estimated residual.

20. Recall that the events making up the proportions are unlikely to be independent and the proportions themselves are unlikely to be independent. Thus conventional significance tests comparing, for example, the pretest differences in proportions against the posttest differences in proportions (McNemar, 1962, pp. 86-88) would have been technically incorrect and could not have been taken literally. In all fairness, however, almost any reasonable discounting of t values would have suggested a statistically significant treatment effect.

REFERENCES

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. New York: John Wiley.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics*. New York: John Wiley.

Berk, R. A. (1988). The role of subjectivity in criminal justice classification and prediction methods. *Criminal Justice Ethics, 7*, 35-46.

Berk, R. A., & Cooley, T. F. (1987). Errors in forecasting social phenomena. *Climatic Change, 11*, 247-265.

Cinlar, E. (1975). *Introduction to stochastic processes*. Englewood Cliffs, NJ: Prentice-Hall.

Colin, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparison and applications of some estimators and tests. *Journal of Applied Econometrics, 1*, 29-53.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Draper, D. (1987). Rank-based robust analysis of linear models. *Statistical Science, 3*(2), 239-258.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54-77.

Granger, C.W.J., & Newbold, P. (1986). *Forecasting economic times series*. Orlando, FL: Academic Press.

Greenspan, R., Berk, R. A., Feeley, M. M., & Skolnick, J. H. (1988). *Courts, probation, and street crime: Final report on the targeted urban crime narcotics task force*. Berkeley, CA: Center for the Study of Law and Society.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: John Wiley.

Hsiao, C. (1986). *Analysis of panel data*. New York: Cambridge University Press.

Huber, P. J. (1977). *Robust statistical procedures*. Philadelphia: Society for Industrial and Applied Mathematics.

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley.

Koenker, R., & Portnoy, S. (1987). Comment. *Statistical Science, 3*(2), 259-261.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economics Review, 73*, 31-43.

Li, G. (1985). Robust regression. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes*. New York: John Wiley.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. New York: Chapman & Hall.

McNemar, P. (1962). *Psychological statistics* (3rd ed.). New York: John Wiley.

Mosteller, R., Fienberg, S. E., & Rourke, R.E.K. (1985). *Beginning statistics with data analysis*. Reading, MA: Addison-Wesley.

Mosteller, R., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Rosenberger, J. L., & Gasko M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297-338). New York: John Wiley.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: John Wiley.

Stone, M. (1988). Quetelet and the poetry of statistical conjecture. *Chance, 1*, 10-16.

Welsch, R. E. (1980). Regression sensitivity analysis and bounded-influence estimation. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 153-167). New York: Academic Press.

Wu, L. L. (1985). Robust M-estimation of location and regression. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 316-388). San Francisco: Jossey-Bass.

8

AN INTRODUCTION TO BOOTSTRAP METHODS
Examples and Ideas

Robert Stine

The bootstrap is an *approach* to estimating sampling variances, confidence intervals, and other properties of statistics. Just as maximum likelihood refers to an estimation strategy rather than to any specific estimator, bootstrapping is a methodology for *evaluating* statistics based on an appealing paradigm. This paradigm arises from an analogy in which the observed data assume the role of an underlying population. As a result, bootstrap variances, distributions, and confidence intervals are obtained by drawing samples from the sample.

Data analysis seeks answers to questions such as "Does a new drug cure more people than the old one?" or "What factors affect how someone votes in an election?" Statistical answers to such questions require models that characterize the random behavior of observed factors. Estimates of the model arise from observed data and lead to description or inference. The importance of the bootstrap lies in this inferential step: The bootstrap gives standard errors and confidence intervals that are typically better than alternatives that rely on untested assumptions. The flexibility of the bootstrap gives the data analyst the freedom to choose statistics whose standard errors would otherwise be difficult to measure. The bootstrap offers reliability and brings new insights to some of the difficult problems of data analysis.