MASTER COPY #84-5

A WORKING PAPER

FROM

The Center for Sociological Research

Department of Sociology

LELAND STANFORD JUNIOR UNIVERSITY ORGANIZED 1891

STANFORD UNIVERSITY STANFORD, CALIFORNIA 94305

A (NOT SO) QUICK AND DIRTY LOOK AT ROBUST M-ESTIMATION*

Lawrence L. Wu
Stanford University

Working Paper No. 84-5
June 1984

*Research support was provided in part by National Science Foundation Grant SES80-23542. I have benefitted greatly from conversations with David L. Donoho, Bradley Efron, M. Vernon Johns, Iain Johnstone, and David J. Pasta. I wish to thank Ross Boylan and Michael T. Hannan for helpful comments on an earlier version. Camille Marder generously made available data files on youth's wages that she had prepared for other purposes. Nancy B. Tuma deserves special credit, not only for extensive comments on this and other versions, but also for suggesting, indirectly, the title for this paper. The International Institute for Applied Systems Analysis provided a supportive atmosphere for completion of early drafts of this paper.

Many statistics used in the social sciences assume that the data conform to a Gaussian (normal) distribution. But such an assumption is rarely an exact statement. On occasion we may believe that the distribution of the data resemble a symmetric bell-shaped distribution. At other times we suspect that such an assumption is at best a rough approximation. In either case, statistics that are insensitive or *robust* to deviations from the distributional assumptions have considerable appeal. The purpose of this chapter is to explain and illustrate one class of robust statistics, the maximum-likelihood type or *M-estimators.*

How do the usual estimators perform under deviations from the distributional assumptions? A surprising result, due to Tukey (1960), shows that many estimators that have optimal performance for data sampled from a Gaussian distribution perform poorly if slight changes are made to the tails of the sampling distribution. Thus, poor performance of the classical estimators can result from the effects of a few gross errors or outliers even if most of the data conform to a Gaussian distribution. This is particularly true for the classical statistics often used by sociologists—estimators of location and spread such as the mean and variance, and estimation procedures for regression such as least-squares, weighted least-squares, and maximum-likelihood. Moreover, since such unusual observations need only involve a few percent of all observations, they are difficult to detect and can be due to transient or unusual phenomena not typical of the underlying population.

A traditional classification distinguishes between parametric and nonparametric estimators. An estimator is said to be *parametric* if it is designed to have optimal performance for a single parametric family of distributions. An estimator is said to be *nonparametric* or *distribution-free* if it is to be used for all nonparameterized distributions. Robust estimators do not easily fall into either category since they are designed to perform well for a broad range of distributions that are thought to provide plausible models of empirical data. Hence, a robust approach to estimation distinguishes between distributions that are more or less plausible, unlike nonparametric approaches, which treat all distributions equally. Indeed, for many distributions the performance of the better robust estimators can be quite close to the best attainable.

A common alternative to robust statistics is to employ a two-step procedure—first clean the data of unusual observations and then apply the classical estimators. Such a procedure is clearly preferable to a naive application of classical estimators to data containing outliers or gross errors. And occasionally we have strong substantive reasons for rejecting observations, for example, if we observe negative values for data that take only positive values. If the correct values cannot be determined, then it is sensible to reject such observations and apply a classical or robust procedure. But we often lack such definitive substantive guidance. In its absence, analysts who adopt a two-stage procedure typically utilize a rejection rule when estimating a location parameter or one or more of the available regression diagnostics procedures when estimating regression parameters.

Do such two-step procedures perform as well as robust estimators? When estimating a location parameter, substantial evidence suggests they do not (Andrews et al., 1972; Hampel, 1974b; Relles and Rogers, 1977; Huber, 1981; Donoho and Huber, 1983). Hampel (1974b) examined several data-cleaning procedures and found they do not perform as well as robust M-estimators of location. The same is true for more informal rejection procedures. Relles and Rogers (1977) compared the "performance" of the best location estimates of several statisticians to the performance of several M-estimators. In analyzing data containing outliers, the statisticians avoided catastrophic failures but were beaten by the better M-estimators. The conclusion is clear: unless the data are free of outliers and errors, the use of a robust estimator of location is preferable to procedures that throw away potential outliers.

When estimating parameters in a regression equation, there is less consensus concerning the relative merits of robust regression and diagnostic procedures partly because both areas are in a state of rapid development (see, for example, Cook, 1977; Andrews and Pregibon, 1978; Hoaglin and Welsch, 1978; Belsley, Kuh, and Welsch, 1980; Atkinson, 1982; Cook and Weisberg, 1982a,b; Gasko and Donoho, 1982; Krasker and Welsch, 1982; Rousseeuw, 1982; Siegel, 1982; Huber, 1983a). Diagnostic procedures, which typically use measures constructed from residuals and deletion statistics obtained from the OLS estimator, seek to identify aspects of the data that deviate from the assumptions. As such, they differ somewhat from robust methods, which seek estimators that can accommodate and control data that deviate from the assumptions.

The following example, adapted from Rousseeuw (1982), illustrates some basic issues in robust estimation as well as some of the peculiar difficulties that can arise in regression.

For this example I generated 30 "good" observations according to

$$y_i = a + bx_i + \epsilon_i ,$$

where $a = 2$, $b = 1$; x_i was drawn from a uniform distribution on (0,10); and ϵ_i was drawn from a Gaussian distribution with mean 0 and standard deviation 0.20. In addition, 20 "bad" observations were drawn from a bivariate Gaussian distribution with mean (22, −2) and standard deviation 0.60. Figure 1 plots the data and estimated regression lines for the OLS solution and two solutions obtained from a robust M-estimator due to Bell (1980).

[Figure 1 about here]

The data in Figure 1 illustrate a classic example of the effects of so-called "leverage" points. A leverage point x_i is defined as an outlier in the x's that can (potentially) exert a strong influence on the parameter estimates or the predicted value of $\hat{y}_i$ by virtue of its position in the data (Belsley, Kuh, and Welsch, 1980; Cook and Weisberg, 1982a,b). More generally, the problems caused by leverage points can be cast in terms of the breakdown properties of an estimator (Gasko and Donoho, 1982), where the breakdown point of an estimator is, roughly speaking, the proportion of "bad" or "contaminated" observations that the estimator can tolerate before yielding an unreasonable estimate. Clearly, the data in Figure 1 represent an extreme test of any regression procedure since a large proportion (40%) of the data are contaminated.

M-estimators are obtained by minimizing the sum of general functions of the deviations; hence, the mean and ordinary least-squares (OLS) estimators, which minimize the sum of squared deviations, are special cases. Because of their greater generality, M-estimates must usually be obtained from a numerical procedure that starts from some initial estimate and iterates until changes in the numerical values are within some desired accuracy. The two M-estimates in Figure 1 correspond to the solutions obtained by starting the iterations from the OLS estimate, denoted Bell/OLS, and from a robust (high breakdown) repeated median estimate (Siegel, 1982), denoted Bell/RM. (See Appendix 1 for a discussion of high breakdown initial estimates.) Both the OLS estimator ($\hat{a} = 8.37$, $\hat{b} = -0.424$) and the Bell/OLS estimator ($\hat{a} = 8.20$, $\hat{b} = -0.431$) yield similar answers—both are clearly affected by the presence of the outlying cluster of points. In contrast, the Bell/RM estimate (starting estimates $\hat{a} = 3.28$, $\hat{b} = 0.816$) yields $\hat{a} = 2.04$ and $\hat{b} = 0.999$, values quite close to those used to generate the first 30 observations.[1]

How do the standard diagnostic tools perform? In this particular example, they fared poorly and failed to identify the outlying cluster of observations in the data.[2] The difficulty lies in the fact that multiple outliers can mask the effects of one another by increasing the size of residuals for other observations (Andrews, 1979; Belsley, Kuh, and Welsch, 1980; Gasko and Donoho, 1982). This problem does not occur with the M-estimator started from the repeated median estimate, and an inspection of the residuals obtained from this estimator easily identifies the outlying cluster of observations.

What if the cluster of points is centered at $(7, -2)$ instead of $(22, -2)$? Because these points now fall within the range (0,10) on the x-axis, they are no longer outliers in the x's and hence are not leverage points. Nevertheless, they exert a large effect on both the OLS estimator ($\hat{a} = 3.50$, $\hat{b} = -0.0259$) and the M-estimator started from the OLS estimate ($\hat{a} = 3.65$, $\hat{b} = -0.0623$). Similarly, the standard diagnostic procedures fail to identify any of 20 contaminated points. But the M-estimator started from the repeated median estimate ($\hat{a} = 2.29$, $\hat{b} = 0.892$) yields the same estimates as before ($\hat{a} = 2.04$ and $\hat{b} = 0.999$).

As these results demonstrate, a robust estimator can yield different estimates for different initial estimates. This clearly differs from classical estimation procedures like OLS, which always yield one estimate given the data and model. Which estimate should one report when one lacks knowledge of the true underlying structure of the data and is confronted with different estimates? In fact, such a question assumes that the data provide one single indication, an assumption clearly violated by the data in Figure 1 in which 60% of the observations follow a simple linear pattern and 40% of the observations are clustered in a spherical pattern. This suggests that multiple solutions for M-estimators can point to multiple features in the data that may not allow a single simple interpretation. For example, the M-estimator using a robust starting estimate was able to accommodate and control the outlying cluster of observations; hence it could be used to estimate the correct parameter values for the observations that conformed to the linear model. But the existence of multiple solutions also suggests inadequacies in our theory or model since the goal of fitting a straight line to all observations in these data is clearly an inappropriate one.

It should be stressed that the problem of what to do with discrepant data is not a simple statistical matter but rather one that concerns the consequences and substantive interpretation of such data. Sometimes we are interested only in how such unusual observations may affect the conclusions drawn from the data. But questions concerning the interpretation of such observations are equally important—are such observations gross errors (that is, observations sampled with large error) or are they outliers (that is, observations that differ in important substantive ways from the rest of the population)? Since outliers, as opposed to gross errors, often provide considerable insight into the data, they merit careful attention.

Clearly, the data of this example represent a "worst-case" situation since a large proportion of the data are contaminated in an asymmetric way by a tight cluster of leverage points. Standard diagnostic procedures may be adequate to identify leverage points and other unusual observations in practice even though they did not fare well in this particular example. Similarly, although graphical methods such as scatterplots cannot always be relied upon to identify clusters of leverage points in high dimensional multivariate data, the problem is quickly and easily detected in the bivariate scatterplot in Figure 1 (Friedman and Stuetzle, 1982). Nevertheless, this example provides a cautionary tale and demonstrates the potential usefulness of a robust estimator.

The rest of the chapter has seven main sections. Since sociologists may be unfamiliar with the large statistical literature on robustness,[3] Section 1 provides a brief overview of underlying issues. Sections 2–4 focus on the robust estimation of a location parameter. Section 2 defines basic terms pertaining to robust estimation. Section 3 gives a general definition of an M-estimator, discusses the need for auxiliary estimators of scale, describes the connection between ML- and M-estimators, and presents several commonly used M-estimators. Section 4 formalizes several desirable properties and concepts introduced informally in Section 1. Section 5 presents some robust hypothesis tests and confidence intervals. Sections 6 and 7 discuss a simple extension of M-estimators of location to multiparameter regression problems and illustrate such an extension with empirical data.

Although some sections assume some familiarity with probability theory and statistical concepts, many are quite straightforward. Sections 1, 5, 6, and 7 present ideas in a relatively informal manner and require little background. Section 2 and parts of sections 3 and 4 are more technical; the more difficult parts in section 4 have been starred.

The previous example suggested that the classical estimators can behave poorly if the data contain a large proportion of unusual observations. The following example, due to Tukey (1960) and reproduced in many sources (Huber, 1977; Mosteller and Tukey, 1977; Huber, 1981; Iglewicz, 1983), demonstrates how rapidly the optimal performance of some classical estimators can deteriorate under even small deviations from the parametric assumptions.

Consider the distribution

$$F(y) = (1-\epsilon)\Phi(\frac{y-\theta}{\sigma}) + \epsilon\,\Phi(\frac{y-\theta}{c\sigma})\,, \tag{1}$$

where $\Phi(\cdot)$ denotes the standard Gaussian distribution

$$\Phi(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} \mathrm{e}^{-\nu^2/2}\,\mathrm{d}\nu\,; \tag{2}$$

ϵ is a small positive number, c is some positive number, and ν is a dummy of integration. The distribution $F(y)$ is commonly called an ϵ-contaminated Gaussian distribution at scale c. (I use y instead of x to emphasize that primary interest concerns deviations from distributional assumptions about the dependent variable, not the independent variables.) It represents a mixture of Gaussian distributions since observations are sampled according to $\text{Gau}(\theta,\sigma^2)$ with probability $1-\epsilon$ and $\text{Gau}(\theta,c^2\sigma^2)$ with probability ϵ, where $\text{Gau}(\theta,\sigma^2)$ denotes a Gaussian distribution with mean θ and variance σ^2.

Note that $F(y)$ resembles a standard Gaussian distribution in the center but has heavier tails, that is, greater mass for large values of $\pm y$. A contaminated Gaussian is often plausible since it models a sample in which most observations are "good" but a small proportion are "bad", that is, sampled with greater variance. Such behavior could occur if a few observations are sampled with gross errors or if the population contains a small proportion of outliers.[4]

Two common statistics for spread are the mean square deviation

$$s_n = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right]^{1/2}$$

and the mean absolute deviation

$$d_n = \frac{1}{n}\sum_{i=1}^{n} |y_i - \overline{y}| \,.$$

A classical result, due to Fisher (1920), states that s_n is about 12% more efficient than d_n under sampling from a Gaussian distribution. That is, under Gaussian sampling, s_n with $n = 88$ gives as concentrated an estimator of spread as d_n with $n = 100$.

What happens if the underlying distribution differs from a Gaussian distribution? In particular, is s_n always more efficient than d_n? There is a slight technical problem since s_n and d_n measure different things. For Gaussian samples, s_n converges to the population standard deviation σ, while d_n converges to $\sqrt{2\pi}\,\sigma$, which is about $.80\sigma$. To compare the efficiencies of s_n and d_n, Tukey (1960; see also Huber, 1977, 1981) used the asymptotic relative efficiency, which takes into account these differences between s_n and d_n

$$\text{ARE} = \lim_{n\to\infty} \frac{\text{Var}(s_n)/[\text{E}(s_n)]^2}{\text{Var}(d_n)/[\text{E}(d_n)]^2} \,.$$

Suppose $c = 3$ in expression (1), that is, we sample from the ϵ-contaminated Gaussian at scale 3. Then (Huber, 1977, 1981)

$$\text{ARE}(\epsilon) = \frac{1}{4}\left[\frac{3(1+80\epsilon)}{(1+8\epsilon)^2} - 1\right] \Big/ \left[\frac{\pi}{2}\,\frac{(1+8\epsilon)}{(1+2\epsilon)^2} - 1\right] . \tag{3}$$

Table 1 gives some values for the expression in (3).

[Table 1 about here]

The results are striking. For any level of contamination between .002 and .50, the mean absolute deviation d_n is a better estimator than the mean square deviation s_n.[5] In the most extreme case ($\epsilon = .05$), d_n is twice as efficient as s_n. In return, we give up 12% efficiency if the data conform *exactly* to a Gaussian distribution, that is, if $\epsilon = 0$. Hence, the use of d_n over s_n entails a cost of some efficiency (sometimes called the "premium," see Anscombe, 1960) if the assumed distribution holds exactly in return for protection ("insurance") against distributions that deviate from the assumed distribution.

This example helps motivate several desirable properties of robust statistics (see Section 4 for a more formal treatment).

1. A robust estimator should satisfy standard statistical properties of consistency, unbiasedness for symmetric distributions, asymptotic normality, and equivariance under translation and scale transformations. Roughly speaking, unbiasedness and consistency require that the estimator yield the correct value θ for the distribution $F(y;\theta)$, where θ is the unknown parameter characterizing F. Asymptotic normality requires that the distribution of $\sqrt{n}\,(\hat{\theta} - \theta)$ be Gaussian as $n \to \infty$, whatever the underlying distribution $F(y;\theta)$. This property allows the construction of simple tests of hypotheses and confidence intervals based on the Student t-distribution. Lastly, translation and scale equivariance require that the estimator be equivariant under location and scale transformations—adding or multiplying all observations by a constant should change the estimator in the same way.
2. The value of a robust estimate should change only slightly for small deviations of the actual distribution from the assumed distribution. Such deviations might be either large changes in a small fraction of the sample (for example, gross errors as in the example above) or small changes in a large fraction of the sample (for example, rounding errors in the data, errors due to the finite number of significant digits in the data, and so forth). Such an estimator is said to be *resistant.*
3. The value of a robust estimate should not change drastically even for large deviations of the actual distribution from the assumed distribution. Large deviations include qualitative changes in the shape of the distribution, for example, large departures from symmetry, as in the example in Figure 1. Such an estimator is said to have a *high breakdown point.*
4. A robust estimator should be efficient for distributions that plausibly model empirical data. Typically, one requires that a robust estimator have high efficiency for distributions that resemble the Gaussian distribution in the center but differ in the tails. Such an estimator is said to be *robust of efficiency.*

Lastly, since most robust estimators satisfy the first three criteria and many of these are efficient for a wide range of distributions, it seems reasonable to require the following.

5. Robust estimators should be *practical.* That is, estimators should be flexible (for example, readily generalized to regression), easy to use and describe, reasonable in

cost, and suitable for the sample sizes that sociologists usually encounter.

The last criterion motivates my focus on M-estimators. Several classes of robust estimators have been studied intensively (see, for example, Huber, 1981): *M-estimators*, which are maximum-likelihood-type estimators; *L-estimators*, which are linear combinations of order statistics; and *R-estimators*, which are derived from rank-order tests. More recently Johns (1979) has proposed a fourth class of robust statistics: *P-estimators*, which are analogues of Pitman estimators. P-estimators have excellent properties and extend naturally to multiparameter problems like regression. However, they require multiple integrals for multiparameter problems, which must be done by numerical integration. L-estimators are attractive in one-parameter problems but do not extend easily to regression. R-estimators and M-estimators extend naturally to regression, but M-estimators are simpler, more flexible, and better understood. M-estimators also appear to have slightly better statistical properties than R-estimators.

Efficiency is often important in practice since it provides a useful measure of how closely an estimator estimates the unknown parameter θ. When the data roughly resemble a bell-shape curve but contain outliers or gross errors, classical estimators like the mean are usually less efficient than nonparametric estimators like the median, which are in turn usually less efficient than the robust M-estimators.[6]

The literature on robustness typically assesses either asymptotic or small ($n \leq 20$) sample performance. Empirical evidence suggests that the performance of robust estimators for $n = 40$ is close to their asymptotic (that is, $n = \infty$) performance. Since the samples that sociologists encounter are often large ($n \geq 40$), the asymptotic performance of M-estimators is likely to predict their performance well for many sociological applications.

2. *DEFINITIONS*

Basic Terms. This section briefly reviews some necessary statistical notions. Let θ be the unknown parameter characterizing the distribution $F(y;\theta)$. A statistic $T_n(Y_1,\ldots,Y_n)$, that is, a known function of random variables $Y_1,\ldots,Y_n$, is an *estimator* of θ if the value of T_n is used to estimate θ. The *estimate* $\hat{T}_n$ refers to the numerical value of T_n obtained from a particular sample $y_1,\ldots,y_n$. (Throughout this chapter I denote a random variable by an upper-case letter and its realization by a lower-case letter.)

The *empirical distribution function* F_n for $Y_1, \ldots, Y_n$ is defined by

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \xi_{y_i}(y) \,, \tag{4}$$

where $\xi_{y_i}(y)$ is the indicator function

$$\xi_{y_i}(y) = \begin{cases} 1, & \text{if } y \geq y_i \,; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Hence, F_n is a step function with jumps of $1/n$ at the observed values y_i.

The estimators considered throughout this chapter may be regarded as functions that depend on the sample only through F_n. Such estimators, called *functional estimators*, are denoted

$$T_n \equiv T(F_n) = T_n(F_n)$$

The above notation emphasizes that the function T does not vary with the sample size n although a more general statistic $T_n(F_n)$ could.

The *likelihood function* $\mathcal{L}$ of n random variables $Y_1, \ldots, Y_n$ is defined as the joint density of the random variables, where the likelihood is considered to be a function of the unknown parameter θ. Let $Y_1, \ldots, Y_n$ be independently and identically distributed (i.i.d.) according to $F(y; \theta)$. Then the likelihood function is given by

$$\mathcal{L}_F(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \theta) \,, \tag{6}$$

where $f(y; \theta)$ is the density corresponding to $F(y; \theta)$.

Maximum-Likelihood Estimators. The *maximum-likelihood* or *ML*-estimator T_n is given by the value of θ that maximizes the likelihood function in (6) or, equivalently, that minimizes

$$-\log \mathcal{L} = \sum_{i=1}^{n} \rho(y_i; \theta) \,, \tag{7}$$

where

$$\rho(y; \theta) \equiv -\log f(y; \theta) \,. \tag{8}$$

Under mild conditions ML-estimators can be shown to have minimum variance among unbiased estimators when the underlying distribution F is known. The function $\rho(y; \theta)$ can often be expressed as a function that depends on y and θ only through $y-\theta$; for the present I retain the more general notation.

If the function $\rho(y;\theta) \equiv -\log f(y;\theta)$ is sufficiently regular, for example, if it is differentiable and convex, then the ML-estimator T_n is the value of θ that is a root of

$$\sum_{i=1}^{n} \psi(y_i;\theta) = 0\,, \tag{9}$$

where

$$\psi(y;\theta) \equiv -\frac{\partial}{\partial\theta}\rho(y;\theta) = \frac{\partial}{\partial\theta}\log f(y;\theta)\,. \tag{10}$$

Convexity guarantees that the solution of expression (9) for T_n is unique. Otherwise several values of T_n, corresponding to local minima or maxima of $\mathcal{L}$, may satisfy (9).

3. *M-ESTIMATORS OF LOCATION*

This section presents a basic outline of M-estimators of location. Goodall (1983) and Huber (1977, 1981) give similar but more systematic treatments of the material covered in this and the subsequent section. I begin with a general definition of an M-estimator, which is defined in a manner analogous to the definition of an ML-estimator.

Definition of an M-estimator. An *M-estimator* T_n is given by the value of θ that minimizes

$$\sum_{i=1}^{n} \rho(y_i;\theta)\,, \tag{11}$$

or is a root of

$$\sum_{i=1}^{n} \psi(y_i;\theta) = 0\,, \tag{12}$$

where ρ is an arbitrary function and $\psi \equiv -\partial\rho/\partial\theta$. Thus, it is sufficient to specify the function ρ or ψ to define an M-estimator, a fact exploited extensively throughout this chapter. Note that ρ and ψ may be more general functions than the corresponding ρ and ψ for ML-estimators, which are derived from the assumed parametric density f. Hence, ML-estimators are special cases of the more general M-estimators.

Location and Scale Equivariance. An estimator T is *scale equivariant* if $T(ay) = aT(y)$, that is, if all observations are multiplied by the constant a, the estimate T is multiplied by the same constant. An estimator T is *location equivariant* if $T(y+b) = T(y)+b$, that is, if all observations are shifted by a value b, the estimate T shifts by the same constant. Classical estimators, such as the mean and median, have the property that $T(ay+b) = aT(y)+b$ and thus are both location and scale equivariant.

As noted earlier, $\rho(y;\theta)$ and $\psi(y;\theta)$ can often be expressed as functions that depend on y and θ only through $y-\theta$. A simple argument shows that M-estimators defined by $\rho(y-\theta)$ or $\psi(y-\theta)$ are location equivariant but need not be scale equivariant.

To make M-estimators scale equivariant it is necessary to introduce an *auxiliary estimator of scale* $S_n \equiv S(F_n)$ into the definitions of the M-estimator. Let

$$u_i = \frac{y_i - \theta}{kS_n}\,, \tag{13}$$

where S_n is a suitable auxiliary estimator of scale and k is a "tuning constant" that can be adjusted to "fine-tune" the performance of the estimator. Then the M-estimator T_n is both location and scale equivariant if it is defined as the value of θ minimizing

$$\sum_{i=1}^{n} \rho(u_i) \tag{14}$$

or solving

$$\sum_{i=1}^{n} \psi(u_i) = 0\,, \tag{15}$$

where

$$\psi(u) \equiv \frac{\partial\rho}{\partial u}\,. \tag{16}$$

To date, two of the most successful estimators for auxiliary scale (see, for example, Andrews et al., 1972) are the *median absolute deviation* (MAD)

$$\mathrm{MAD}(y_i) = \mathrm{median}\,|y_i - T_n| \tag{17a}$$

and the interquartile range

$$Q(y_i) = F_n^{-1}(3/4) - F_n^{-1}(1/4)\,,$$

where $F_n^{-1}(\cdot)$ denotes the sample percentiles of $y_i - T_n$. A normalizing constant is sometimes introduced into the expression for the MAD

$$\text{normed MAD}(y_i) = \frac{\mathrm{median}\,|y_i - T_n|}{0.6745}\,. \tag{17b}$$

The factor of 0.6745 makes $S_n = \sigma$ if the data are sampled from a Gaussian distribution with variance σ^2.

Typically the location estimator is of primary interest, and the estimator of auxiliary scale is a "nuisance" parameter that serves to make the location estimator scale equivariant. Thus the estimation of the auxiliary scale parameter presents different problems than the estimation of location. In particular, bias and breakdown appear to be more important criteria than efficiency (Bell. 1980; Huber, 1981).

Estimation. Note that the expression in (15) can be reexpressed as a weighted mean problem. To see this, rewrite (15) as

$$\sum_{i=1}^{n} u_i w_i = 0 ,$$

where the weights w_i depend on the sample according to $w_i = \psi(u_i)/u_i$ for $u_i \neq 0$. Then T_n is the weighted mean given by

$$T_n = \sum_{i=1}^{n} u_i w_i \bigg/ \sum_{i=1}^{n} w_i \ . \tag{18}$$

Hence, to estimate T_n, one can either minimize the expression in (14) using an iterative Newton–Raphson type algorithm or iterate on the expression in (18) until the sequence of estimated values for T_n converges to a desired accuracy. In either case, one must start the iterations from some initial estimate of location, which is typically chosen to be the median. Then one can use the location estimate to compute the auxiliary estimate of scale and vice versa.[7]

Henceforth I restrict attention to translation and scale equivariant estimates of location under the assumption of a symmetric distribution F. The assumption of a symmetric distribution is strong but often plausible, particularly if suitable transformations of the data are allowed. Note, moreover, that an estimator with a high breakdown point can tolerate large asymmetric departures from the distributional assumptions. Under the assumption of symmetry, the center about which the density is symmetric provides a natural population parameter for θ, unlike the more general case of an asymmetric distribution, where the definition of a natural location parameter is problematic.

Desirable Shapes for ρ and ψ. Below I derive ρ and ψ for the ML-estimators for the Gaussian, logistic, double exponential, and Cauchy distributions. As noted earlier, ML-estimators have minimum variance among unbiased estimators if the parametric distribution F is known; hence the fully efficient M-estimator *is* the ML-estimator if the data in fact conform to F. But it is important to stress that ML-estimators need not be—and often are not—robust of efficiency for other distributions. Nevertheless, calculating such expressions for ρ and ψ serves a useful heuristic purpose by showing how ρ and ψ for the ML-estimators change with different distributional shapes. In particular, the shape of ψ provides important insights that help guide the construction of M-estimators possessing higher efficiency for a wide range of distributions.

Let the sampled observations be i.i.d. according to the known density f. For simplicity, I assume that the scale of f is known and fixed such that $f(\theta) = 1/\sqrt{2\pi}$. Then ρ and ψ for ML-estimators are given by (8) and (10). Table 2 gives expressions for the normalized densities for the Gaussian, logistic, double exponential, and Cauchy distributions as well as the functions ρ and ψ corresponding to the ML-estimators for each distribution. Note that ρ and ψ for ML-estimators of the Gaussian and double exponential distributions define two classical estimators of location, the mean and median, respectively; ρ and ψ for the ML-estimators for the logistic and Cauchy distributions do not correspond to any familiar estimators.

[Table 2 about here]

To fix ideas, I calculate ρ and ψ explicitly for the ML-estimator for the double exponential distribution. By (8) and (10), $\rho \equiv -\log f(y;\theta)$ and $\psi \equiv -\partial\rho/\partial\theta = \partial[\log f(y;\theta)]/\partial\theta$. The normalized density function for the double exponential distribution is given by

$$f(y;\theta) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\sqrt{2/\pi}\,|y-\theta|} .$$

Taking logs and changing signs yields

$$-\log f(y;\theta) = \frac{1}{2}\log 2\pi + \sqrt{\frac{2}{\pi}}\,|y-\theta| . \tag{19}$$

Hence

$$\rho(u) = \frac{1}{2}\log 2\pi + \sqrt{\frac{2}{\pi}}\,|u| ,$$

where $u \equiv y - \theta$. Note that $\rho(u)$ is proportional to $|u|$ up to additive and multiplicative constants that arise from the normalizing constants in the density f.

Differentiating (19) with respect to θ gives

$$-\frac{\partial}{\partial\theta}\log f(y;\theta) = \sqrt{\frac{2}{\pi}}\,\frac{\partial}{\partial\theta}\,|y-\theta|\ .$$

Then by (10)

$$\psi(u) = \sqrt{\frac{2}{\pi}}\,\mathrm{sign}(u)\ ,$$

where sign(u) is defined by[8]

$$\mathrm{sign}(u) = \begin{cases} 1, & \text{if } u \geq 0\,; \\ -1, & \text{otherwise.} \end{cases}$$

Figure 2 illustrates the densities in Table 2, which are symmetric about $\theta = 0$. The left-hand side of Figure 2 illustrates the central shapes of the densities and plots $f(y)$ for $-3 \leq y \leq 0$. To compare the extreme tail behavior of the densities, the right-hand side of Figure 2 plots $f(y)$ for $3 \leq y \leq 12$; for convenience, the scale of $f(y)$ is increased by a factor of 10 in this range. Note that the logistic, double exponential, and Cauchy distributions have progressively heavier tails than the Gaussian distribution, which fall off as $\exp(-y^2)$. In particular, the tails of the double exponential distribution fall off as $\exp(-|y|)$ while those for the Cauchy fall off as $1/y^2$. Note also the distinctive central shapes of the double exponential and Cauchy distributions, which are more sharply peaked than the Gaussian and logistic distributions.

[Figure 2 about here]

Figure 3, which illustrates $\psi(u) = \partial[\log f(y;\theta)]/\partial\theta$ for the four ML-estimators, demonstrates these same characteristics somewhat more vividly. The similar central shapes of Gaussian and logistic distributions correspond to shapes of ψ that are linear or close to linear near the origin. The sharp central peaks of the double exponential and Cauchy distributions correspond to large slopes of ψ near the origin (ψ for the median, which is the ML-estimator for the double exponential distribution, is actually discontinuous at the origin). The shapes of ψ also illustrate the differences in the tails of the distributions. For large $\pm u$, $\psi(u)$ is large in absolute value for lighter tailed distributions but small in absolute value for heavier tailed distributions. In particular, ψ is unbounded for the Gaussian distribution but bounded for the other distributions. In the case of the Cauchy distribution, which has the heaviest tails, $\psi(u)$ even "redescends," that is, $\lim \psi(u) = 0$ as $u \rightarrow \pm\infty$.

[Figure 3 about here]

Clearly, the functions ρ and ψ for ML-estimators tell little more than the density f; the function $\rho \equiv -\log f(y;\theta)$ is simply $f(y;\theta)$ on a log scale while $\psi = \partial[\log f(y;\theta)]/\partial\theta$ gives the rate of change for $\log f(y;\theta)$. But since the functions ρ and ψ for M-estimators need not be fixed functions of any density f, we might seek to combine features of different ρ and ψ functions based on the insights gained from examining ρ and ψ for the ML-estimators. Hence the shapes of ρ and ψ for the ML-estimators provide a heuristic tool for constructing more general M-estimators. For example, Figure 3 suggests that a bounded or redescending ψ function may be desirable if the data contain extreme outliers. Similarly, Figure 3 suggests that if ψ is approximately linear near the origin, then the resulting M-estimator is likely to have good performance for distributions resembling a Gaussian distribution in the center.

Some Common M-estimators of Location. This section introduces several common M-estimators of location and motivates these estimators using the heuristic insights gained by examining the ψ in Figure 3. Table 3 gives expressions for ρ, ψ, and $\phi \equiv d\psi/du$ for seven estimators of location: two classical estimators, the mean and median, and five M-estimators proposed by Huber, Hampel, Andrews, Tukey, and Bell (Huber, 1964; Andrews et al., 1972; Beaton and Tukey, 1974; Bell, 1980).

[Table 3 about here]

Figure 4 illustrates ρ, ψ, and ϕ for the mean. To illustrate how one obtains M-estimates, Figure 4 includes three hypothetical points, $\hat{u}_1 = -3$, $\hat{u}_2 = 1$, and $\hat{u}_3 = 2$, corresponding to the deviations $u_i = y_i - \theta$ of three observations y_1, y_2, and y_3 obtained by setting θ equal to the sample mean. (Note that $u_i = y_i - \theta$ for the mean and median since neither requires a tuning constant or an auxiliary estimator of scale.) By definition, the mean is the value of θ minimizing $\sum \rho(u_i)$, where $\rho(u_i)$ is a function proportional to the sum of squared deviations. Hence, the sum of the $\rho(u_i)$ for the three points in Figure 4 can be shown to be the minimum over all possible values of θ. Equivalently, the mean can be defined by the value of θ solving $\sum \psi(u_i) = 0$; hence, the $\psi(u_i)$ in Figure 4 sum to zero.

[Figure 4 about here]

Figure 4 helps to explain why the mean behaves poorly for data containing outliers or gross errors. The plot of $\rho(u)$ shows that ρ is a rapidly increasing function of u. Hence, adding or deleting observations with large positive or negative u can exert a large effect on the estimate $\widehat{T}_n$, so that T_n is a highly nonresistant estimator. Expressing the M-estimator as a weighted mean leads to the same conclusion. By (18) the M-estimator can be expressed as a weighted mean with weights $\psi(u)/u$; hence, the mean assigns a weight of 1 to all observations, including observations far from the bulk of the data. Similarly, because $\psi(u) = u = y - \theta$ is unbounded in u, one observation placed at $y = \pm\infty$ moves the estimated value of the mean to $\pm\infty$. Thus, one sufficiently aberrant value y can cause T_n to have any arbitrary value.

Figures 5–7 illustrates ρ, ψ, and ϕ, respectively, for the remaining estimators in Table 3. Since $\psi(u) = \text{sign}(u)$ for the median, only information on the sign of $u = y - \theta$ is used to obtain the location estimate $\widehat{T}_n$. That is, given two positive values u_1 and u_2 with $u_1 < u_2$, $\psi(u_1)$ is identical in value to $\psi(u_2)$; moreover, letting $u_2 \rightarrow \infty$ has no effect on the estimate $\widehat{T}_n$, demonstrating the resistance of the median to outlying observations. The discontinuity of ψ at $u = 0$ is reflected by $\phi(u) = \delta(u)$, where $\delta(u)$ denotes the Dirac delta function, which is, loosely, a function with an infinite spike for $u = 0$ and identically zero otherwise.[9]

[Figures 5–7 about here]

The Huber M-estimator combines features of the mean and median, providing better protection against extreme outliers than the mean while giving better efficiency for the Gaussian distribution than the median. The shape of ψ, illustrated in Figure 6, is linear in u for $|u| \leq c$ and constant for $|u| > c$. Similarly, $\rho(u)$, shown in Figure 5, is a convex function proportional to u^2 for $|u| \leq c$ and linear in u for $|u| > c$. This allows the Huber M-estimator to act like the mean for centrally located observations and like the median for observations far removed from the bulk of the data.

The tuning constant c and the auxiliary estimator of scale S_n jointly serve to rescale the y_i. Observations less than cS_n units from T_n fall on the linear part of ψ (strictly convex part of ρ) while observations greater than cS_n units from T_n fall on the constant part of ψ (linear part of ρ). Hence the tuning constant c allows one to adjust the performance of the estimator to achieve a desired efficiency for a particular distribution.[10] For example, larger values of the tuning constant c correspond to higher efficiencies of the Huber M-estimator for the Gaussian distribution since the estimator increasingly resembles the mean as $c \longrightarrow \infty$; similarly, the estimator increasingly resembles the median as $c \longrightarrow 0$. This shows that the definition of each M-estimator in Table 3 refers more properly to a family of M-estimators since different values of the tuning constants produce estimators that differ in performance but have similar overall characteristics as determined by the qualitative shapes for ρ and ψ.

Huber (1964) gives a more technical motivation for his M-estimator and shows that it is the ML-estimator for a "least informative" distribution. Consider

$$F(y) = (1 - \epsilon)\Phi(y) + \epsilon G(y) , \tag{20}$$

where $\Phi(y)$ is Gaussian and $G(y)$ is symmetric. If $G(y)$ is chosen so that $F(y)$ has tails that fall off as $\exp(-|y|)$, then $F(y)$ can be shown to have minimum Fisher information for all F in (20) with symmetric G. A standard theorem states that the inverse of the Fisher information is the Cramér-Rao lower bound on the asymptotic variance of all unbiased estimators. Hence, minimizing the Fisher information results in maximum variance for estimators of (20). Since ML-estimators have minimum variance among unbiased estimators, it follows that the Huber M-estimator minimizes the maximum variance for all distributions in (20) with symmetric G.

The Hampel M-estimator may be regarded as a refinement of the Huber M-estimator. Like the Huber estimator, ψ is linear for values of $|u| < a$ and constant for u between a and b. However, Figure 6 shows that ψ redescends linearly towards zero for $b \leq |u| < c$ and is identically zero for $|u| > c$. This permits the estimator to downweight outlying observations progressively, providing additional protection against extremely aberrant observations. Similar performance can be obtained by rejecting extreme outliers and applying the Huber M-estimator to the remaining data except that the Hampel M-estimator downweights such observations in a continuous manner, unlike procedures that identify and reject potential outliers. Because ψ redescends, ρ for the Hampel estimator illustrated in Figure 5 is not a convex function, as was true for the mean, median, and Huber estimators, but is constant for $|u| > c$. This may present technical difficulties since many values of T_n, corresponding to different roots of $\sum \psi(u)$ or to local minima or maxima of $\sum \rho(u)$, may satisfy (14) or

(15).

The sine M-estimator, proposed by Andrews, is defined by a ψ function that consists of one oscillation of the trigonometric sine function; hence, ψ redescends and is identically zero for sufficiently large $\pm u$. Figure 5 shows that the $\rho(u)$ for the Hampel and sine estimators have roughly similar shapes; Figures 6 and 7, illustrating $\psi(u)$ and $\phi(u)$, highlight the differences between the two estimators. One advantage of the sine estimator over the Hampel estimator is that it requires only one tuning constant while the Hampel estimator requires three. A disadvantage is that ψ for the sine estimator ascends and redescends at equal rates, unlike the Hampel ψ function, which is typically adjusted to redescend at a slower rate than it ascends. In particular, this property for the sine ψ function can result in an inconsistent estimator for certain multimodal densities (see the discussion on consistency in Section 4).

One unappealing aspect of the Huber, Hampel and sine M-estimators is that ψ changes slope abruptly for certain values of u, resulting in discontinuities for the ϕ in Figure 7. The ψ function for Tukey's bisquare is similar in shape to ψ for Andrews' sine but changes slope more smoothly. This allows ϕ to be continuous and ψ to redescend at a slightly slower rate than it ascends. It should also be noted that the bisquare enjoys greater popularity than other M-estimators and has received the most empirical study.

The shape of ψ for the Bell M-estimator roughly resembles the shapes of ψ for the sine and bisquare M-estimators. However, ψ for the Bell M-estimator redescends only asymptotically, unlike the "hard" redescending ψ functions of Hampel, Andrews, and Tukey, which are identically zero for sufficiently large $\pm u$. In particular, ψ possesses an infinite number of higher order derivatives that are everywhere continuous, unlike the four M-estimators previously considered. This accounts for the smoother overall shape for ψ and eliminates the abrupt changes in slope for ϕ characteristic of the other M-estimators in Table 3.

Note that the sine, bisquare, and Bell M-estimators have a single tuning constant k that can be adjusted to achieve different efficiencies for different distributions. Since the tuning constant k appears in the denominator of $u = (y - \theta)/kS$, the effect of a larger value of k forces a larger proportion of the u_i's onto the central portion of the ψ function, where the shape of the M-estimator is close to linear, resulting in better efficiency for the Gaussian distribution. Similarly, smaller values of the tuning constants force more of the u_i's onto the tails of the ψ function, resulting in better performance for heavy tailed distributions.

To summarize the previous discussion about the M-estimators in Table 3, an examination of ρ and ψ for the mean helps to suggest why this classical estimator of location lacks resistance for data containing outliers. Figure 4 shows that ρ is a rapidly increasing function of u and that ψ is unbounded in u; hence the estimator is adversely affected by extreme outliers or gross errors. Similarly, an examination of the ML-estimators for various parametric distributions suggest that ρ and ψ for the mean correspond to the light tails of the Gaussian distribution, which fall off extremely rapidly as $\exp(-y^2)$.

The remaining estimators in Table 3 have bounded ψ functions and so offer much greater resistance against outliers. The median has a ψ function that is bounded but discontinuous at the origin. An examination of the double exponential distribution, for which the median is the ML-estimator, suggests that the discontinuity of ψ corresponds to the unusual central peak of the double exponential density. The Huber M-estimator provides a compromise between the mean and median since ψ is continuous and linear for u near the origin but constant for large $\pm u$. It can also be motivated as the ML-estimator for the least-informative distribution, which is Gaussian in the center but has tails that fall off as $\exp(-|y|)$, like those for the double exponential distribution.

The other M-estimators are defined by continuous ψ functions that are approximately linear near the origin but redescend for large $\pm y$. Hence these estimators offer greater protection against distributions with tails that fall off more slowly than $\exp(-|y|)$. The Hampel M-estimator is similar to the Huber estimator but adds an additional linear segment that allows ψ to redescend; because of this, it requires 3 tuning constants. Andrews' sine M-estimator is also defined by a redescending ψ function but is simpler in definition. Tukey's bisquare M-estimator is similar to the sine M-estimator but has a smoother ψ function for u close to ± 1, which permits ϕ to be continuous. The Bell estimator is defined by a ψ function that redescends asymptotically, resulting in an even smoother shape for ψ than the bisquare.

4. *CRITERIA FOR ROBUST STATISTICS*

This section formalizes several desirable properties first introduced in Section 1. The more technical sections have been starred and some readers may wish to skim these sections on a first reading.

**Standard Statistical Criteria.* An estimator $T_n = T(F_n)$ is *unbiased* if and only if (iff)

$$\mathrm{E}_F(T_n) = \theta\,, \tag{21}$$

where θ is the location parameter characterizing the distribution $F(y;\theta)$. If $F(y;\theta)$ is symmetric about θ and ρ is an even function (equivalently, ψ is odd or ϕ is even), then T_n is unbiased.

If F_n converges to F, for example, if F_n is the empirical distribution function of observations sampled according to F, then $T_n \equiv T(F_n)$ is *consistent* iff for all $\epsilon > 0$

$$\lim_{n\to\infty} \Pr(|\,T_n - \theta\,| > \epsilon) = 0\,. \tag{22}$$

That is, consistency requires that T_n converge in probability to the population value θ as $n \to \infty$. In contrast, bias simply requires that the mean value of T_n equal θ.

Regularity conditions for consistency of M-estimators are somewhat stringent, and some M-estimators can be shown to be inconsistent under certain special conditions even if F is symmetric. Roughly speaking, if $\rho(u)$ is convex or if the underlying density f is strongly unimodal, the M-estimator is consistent (Huber, 1981; Freedman and Diaconis, 1982). If the density f is multimodal and ρ is not convex, the M-estimator can be inconsistent.[11]

If the estimator is consistent, then under certain regularity conditions (Huber, 1981) $\sqrt{n}\,(T_n - \theta)$ is asymptotically distributed according to $\mathrm{Gau}(0, A^2(T;F,S))$, where S is the estimator for the auxiliary scale parameter and $A^2(T;F,S)$—see expression (28) below—denotes the variance of the estimator T.

**Influence Curve.* One way of evaluating an estimator is to see how it is affected by one additional observation y in a very large sample. This idea leads to the *influence curve* (Hampel, 1974a)

$$\mathrm{IC}(y;F,T,S) = \lim_{\epsilon\to 0} \frac{T(F_\epsilon(\nu)) - T(F(\nu))}{\epsilon}\,, \tag{23}$$

where

$$F_\epsilon(\nu) = (1-\epsilon)F(\nu) + \epsilon\,\xi_y(\nu)\,, \tag{24}$$

$F(\nu)$ is the underlying distribution, ϵ is a small positive number, and ξ is the indicator function defined in (5). (I have replaced the usual $F(y)$ by $F(\nu)$ for notational clarity.) The expression in (24) states that $F_\epsilon(\nu)$ and $F(\nu)$ differ only by the presence of a point mass of size ϵ at $\nu = y$. Since $F_\epsilon(\nu)$ approaches $F(\nu)$ as $\epsilon \rightarrow 0$, the influence curve may be viewed as the derivative of T_n with respect to ϵ. It thus provides a useful qualitative picture of the asymptotic behavior of T under infinitesimal changes in the underlying distribution.[12]

Under certain regularity conditions (Huber, 1981), the influence curve for M-estimators is

$$\mathrm{IC}(y; F, T, S) = \frac{kS(F)\,\psi\Big(\dfrac{y-T(F)}{kS(F)}\Big)}{\displaystyle\int \phi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\,\mathrm{d}F(\nu)}\,, \tag{25}$$

where F is symmetric and ψ is an odd function. (Appendix 2 presents a derivation for 25.) For example, the influence curve for the mean, which does not require an auxiliary estimator S for scale, is

$$\mathrm{IC}(y; F, T) = y\,.$$

Similarly, the influence curve for the median is (Huber, 1981; Goodall, 1983)

$$\mathrm{IC}(y; F, T) = \frac{\mathrm{sign}(y)}{2f(F^{-1}(1/2))}\,.$$

Note that the expression in (25) varies with F only through $S(F)$ and the integral in the denominator. Thus, evaluating (25) for different F produces influence curves that have identical qualitative shapes but different magnitudes. This provides an asymptotic justification for choosing a shape for ψ that compromises between the shapes for ψ corresponding to ML-estimators for different distributions F.

Resistance. A robust estimator is said to be *resistant* if it is insensitive to large changes in a few observations or small changes in many observations. This property is of considerable appeal since social scientific data often contain three types of observations that can cause the classical estimators to lack resistance: gross errors (for example, observations that have been incorrectly measured or coded), outliers (for example, observations that have large substantive differences from the rest of the population), and small errors (for example, observations that are close but not equal to their true values because of rounding errors).

As noted in the previous section, ψ provides useful information for describing the qualitative behavior of M-estimators under large changes to a few observations. For example, the boundedness of ψ is necessary to insure the insensitivity to large changes in a few observations. Hence, one aberrant observation can completely determine the value of T_n if ψ is unbounded.

The requirement that an estimator be insensitive to *small* changes in many observations implies that ψ be continuous, since many observations occurring at a discontinuity of ψ could change the value of T_n sharply. For example, many strategically placed rounding errors can cause instability in the estimated value of the median but not of the mean.

Consequently, an M-estimator is said to be resistant *iff ψ is bounded and continuous.* This definition provides a remarkably simple rule by which to assess the resistance of an estimator. For example, the mean and median lack resistance since ψ is unbounded for the mean and discontinuous for the median. All other M-estimators discussed in Section 3 are easily seen to be resistant. In practice, the discontinuity of ψ for the median presents fewer difficulties than does the unboundedness of ψ for the mean since, for finite n, the discontinuity of ψ causes sharp but bounded fluctuations of the value of T_n.[13] Hence the median is commonly regarded as resistant despite the discontinuity of ψ at the origin.

Breakdown Point. The *breakdown point* ϵ^* of an estimator is, roughly speaking, the smallest proportion of the sample that can be arbitrarily corrupted before the estimator produces a large aberrant value. An aberrant value might be $\pm\infty$ for a location estimator or ± 1 for a correlation estimator. While breakdown provides a somewhat crude measure of robustness, it is nevertheless extremely useful for examining the effects of large departures from the distributional assumptions, in particular, large asymmetric corruptions of the sample. Because data often have marked asymmetries, breakdown may well be the criterion of greatest practical importance.

The notion of breakdown is originally due to Hampel (1974a), but I present a more recent finite sample definition due to Donoho (1982; see also Donoho and Huber, 1983; Huber, 1983b) that need not refer to any sampling distribution F. I consider two types of arbitrarily corrupted samples Y^* of the fixed sample Y. Let $Y = (y_1, \ldots, y_n)$ be a fixed sample of size n and let $Z = (z_1, \ldots, z_m)$ consist of m arbitrary values. Then a sample corrupted by *ϵ-contamination* is defined by the sample $Y^* = Y \cup Z$ consisting of $n + m$ points, where $\epsilon = m/(n + m)$ is the proportion of contaminated points. Similarly, a sample Y^* corrupted by *ϵ-replacement* is defined by replacing a subset of m points of Y by $Z = (z_1, \ldots, z_m)$, where $\epsilon = m/n$ is the proportion of contaminated points. A sample that is corrupted either by ϵ-contamination or ϵ-replacement is called an *ϵ-corrupted sample.*

Given an ϵ-corrupted sample Y^*, the *maximum bias* of T is defined as

$$b(\epsilon; Y, T) = \sup_{Y^*} |\, T(Y^*) - T(Y)\, | \,, \tag{26}$$

where the supremum is over all ϵ-corrupted samples Y^*. Then the *breakdown point* $\epsilon^*(Y, T)$ for the location estimator T is defined as

$$\epsilon^*(Y, T) = \inf_{\epsilon}\{\, b(\epsilon; Y, T) = \infty \} \,. \tag{27}$$

Note that values of ϵ^* range between 0 and 1. For example, the sample mean has a breakdown point that approaches zero as $n \to \infty$. Adding one sufficiently bad observation to a sample of size n can change the value of the mean by an arbitrary amount; hence $\epsilon^* = 1/(n + 1)$ for samples corrupted by ϵ-replacement. An estimator can also have a breakdown point of 1, for example, the trivial estimator that gives a constant value regardless of the sample. However, the breakdown point of a translation equivariant estimator is at best 1/2. If $\epsilon^* = 1/2$, the translation equivariant estimator cannot distinguish between the original sample Y and the set of contaminating points Z and must break down. The median attains this upper breakdown point of $\epsilon^* = 1/2$ and so has excellent breakdown.

Other M-estimators also have excellent breakdown (Donoho and Huber, 1983). For the M-estimators in Table 3, breakdown is largely determined by the breakdown of the auxiliary estimator for scale and, if ψ redescends, the breakdown of the initial starting estimator. (See Appendix 3 for a derivation due to Huber (1983b) for the breakdown point of redescending M-estimators.) Since the MAD and the interquartile range have breakdown points of 1/2 and 1/4, respectively, a common procedure for estimating a location parameter is to start the iterations from the sample median using the MAD as the estimator for auxiliary scale. If this procedure is used and the tuning constant is chosen for reasonably high efficiency for the Gaussian distribution, then the breakdown point for M-estimators typically exceeds 0.40 or 0.45 (Huber, 1983b).

How large should the breakdown point be? An instructive example concerns two Huber estimators studied in Andrews et al. (1972). The estimators were identical except that one used the MAD for auxiliary scale and had a breakdown point of 1/2, while the other used the interquartile range and had a breakdown point of 1/4. In small sample Monte Carlo simulations, the estimator using the MAD performed substantially better than the estimator using the interquartile range, especially for sampling distributions with heavy tails. Although the two estimators have identical asymptotic properties for all symmetric distributions, the estimator using the MAD appeared to deal better with the random asymmetries that occur in finite sampling from heavy tailed distributions. This suggests that the difference between a breakdown point of 1/2 and 1/4 can lead to a substantial difference in performance.

Efficiency. This section reports small sample ($n = 20$) and asymptotic efficiencies of several estimators for three sampling distributions. It should be noted that high efficiency is not always necessary. If we wish to describe or explore the data, a resistant but relatively inefficient estimator like the median often suffices to summarize the data roughly. In other circumstances nonparametric procedures are appropriate when we are unwilling to make more than weak assumptions. However, more efficient estimators are often needed to obtain precise estimates or to reject inappropriate hypotheses. In such cases efficiency is an important consideration since it lets us assess how closely T_n estimates the unknown parameter θ.

The fourth criterion in Section 1 stated that a robust estimator should have high efficiency for a range of distributions that cover the distributions we may encounter in practice. But what is a plausible range of distributions? As noted earlier, suitable transformations of the data can often make the distribution of the bulk of the data resemble a bell-shape curve, a tendency sometimes called Winsor's principle—that "all distributions are Gaussian in the middle" (Tukey, 1960, p. 457). This suggests the use of a few sampling distributions that resemble the Gaussian distribution in the center but differ in the heaviness of the tails. Following Tukey and others (Beaton and Tukey, 1974; Tukey, 1979b), I examined three distributions: a unit Gaussian, a 5% contaminated Gaussian at scale 10 (see expression 1), and a *slash* distribution defined by a unit Gaussian distribution divided by a uniform (0,1) distribution, which is given formally (Rogers and Tukey, 1972) by

$$f(y) = \frac{1 - e^{-y^2/2}}{y^2 \sqrt{2\pi}} .$$

The Gaussian distribution has comparatively light tails that fall off as $\exp(-y^2/2)$. While such tails are lighter than might be expected for most social scientific data, we may encounter such data on occasion. The contaminated Gaussian distribution has moderately heavy tails. Such a distribution is often plausible because it models a situation in which one small segment of the population is measured with greater error or differs substantively from the remaining population. The slash distribution has extremely heavy, Cauchy-like tails that fall off as $1/y^2$ for large $\pm y$. Since the slash distribution can be obtained by dividing a unit Gaussian deviate by an unit uniform deviate, it can be regarded as a continuous mixture of Gaussian distributions with variances ranging from 1 to ∞. Hence, the slash distribution represents a more radical alternative to the contaminated Gaussian, which is a mixture of two Gaussian populations, and models a population in which variability in the population or measurement error ranges from a fixed lower limit to arbitrarily large values.

For the finite sample variances, sampling from the so-called *one wild Gaussian* (1WG) is used in place of sampling from a 5% contaminated Gaussian at scale 10. For $n = 20$, the 1WG takes 19 observations from a unit Gaussian and 1 observation from a Gau(0, 100). Although the two sampling plans appear similar, the 1WG samples according to a nonprobabilistic rule—it always samples 19 observations from one population and 1 observation from the other population—and so it is not a probability distribution.

The results in Table 1 showed that an estimator that is 100% efficient for one distribution may perform poorly for other distributions. This suggests that in place of optimal efficiency for one distribution, we should instead seek estimators that have "high" efficiency for a wide range of distributions. Tukey (1979a, p. 104) argues that an estimator with 90% efficiency should be regarded as highly efficient:

> "ALL efficiencies between 90% and 100% are NEARLY the SAME for the USER. ... Alternate feedings of bodies of data to 2 statisticians, one of whom uses a 90% efficient estimate, the other using a 100% efficient

> estimate, followed by comparing each's estimates with the corresponding truths, has to involve like 3000 bodies of data before we can prove which is which. Nothing methodological that takes this much data to check is likely to be important."

Tables 4 and 5 report small sample ($n = 20$) and asymptotic relative efficiencies for a number of M-estimators of location for the Gaussian, 1WG, contaminated Gaussian, and slash distributions. The efficiency of an estimator T relative to a reference estimator T_0 for a distribution F is defined by

$$\mathrm{eff}_{T_0}(T, F) = 100 \times \frac{\mathrm{Var}_F(T_0)}{\mathrm{Var}_F(T)} ,$$

where T_0 is chosen to be the estimator with the smallest variance for F. Asymptotic relative efficiencies are obtained from an expression for the asymptotic variance $A^2(T; F, S)$ (see Appendix 4 for a derivation)

$$A^2(T; F, S) = \int [\mathrm{IC}(y; F, T, S)]^2 \mathrm{d}F(y) . \tag{28}$$

Small sample relative efficiencies are taken from the results of Monte Carlo simulations reported in Andrews et al. (1972), Bell (1980), and Goodall (1983).[14]

[Tables 4 and 5 about here]

With the exception of the 1WG, the reference estimators are the Pitman estimators for the finite sample results and the ML-estimators for the asymptotic results. The Pitman and ML-estimators have the smallest finite sample and asymptotic variances, respectively, of all unbiased estimators; the ML-estimator coincides with the Pitman estimator as $n \to \infty$. In the case of the 1WG, neither the Pitman nor ML-estimators are well-defined since the 1WG samples observations in a nonprobabilistic way. Instead, the subsample mean (Kafadar, 1982), a pseudo-estimator, was used as the reference estimator. The subsample mean assumes knowledge of the "wild" observation, that is, the observation sampled according to Gau(0, 100), and computes the mean of the remaining 19 observations.[15]

A glance at Tables 4 and 5 reveals the poor performance of the mean. This estimator is the Pitman and ML-estimator for the Gaussian distribution and so has the best small sample and asymptotic efficiency for Gaussian samples. However, it has only 16–17% efficiency for distributions with moderately heavy tails like the 1WG and contaminated Gaussian and produces extremely variable estimates for the slash distribution, which does not have finite moments. Overall, the mean has the worst efficiency—by substantial margins—of all estimators considered for the 1WG, contaminated Gaussian, and slash distributions.

Other estimators have significantly better performance. For example, the median does better than the mean for the 1WG, contaminated Gaussian, and slash distributions. One gives up 35% efficiency if the data conform exactly to a Gaussian distribution in return for gains of 40–80% in efficiency for heavier tailed distributions. Such a trade-off is often reasonable for exploratory or descriptive analyses, especially in light of the other good properties (high breakdown and resistance) that the median possesses.

A closer examination of Tables 4 and 5 reveals no uniformly best estimator, although some estimators—notably the mean—can be eliminated from consideration. The Huber estimators do well for Gaussian samples but less well for 1WG, contaminated Gaussian, and slash samples. Although the Huber estimators are definitely preferable to the mean and generally perform better than the median, other M-estimators in Tables 4 and 5 have better overall performance.

Some estimators give up a moderate amount of efficiency for the slash distribution in return for better efficiency at the Gaussian distribution, for example, one of the Hampel estimators (a=2.5) and one of the bisquare estimators (k=8.8). Andrews' sine estimator falls into this group but does slightly better for the slash and slightly worse for the Gaussian.

One helpful criterion for assessing efficiency is the *triefficiency* of an estimator (Beaton and Tukey, 1974), defined as the minimum relative efficiency of an estimator for the three sampling distributions. Table 6 reports the four estimators with highest small sample and asymptotic triefficiency. For the $n = 20$ samples, the Bell estimator has a triefficiency of 88.4%, which is the highest small sample figure among estimators considered. A bisquare estimator with tuning constant $k = 6.4$ has the next best triefficiency (86.9%), followed by two Hampel estimators with triefficiencies of 83.0% and 81.8%. The asymptotic results differ slightly. The best estimator is a bisquare estimator ($k = 6$), which has a triefficiency of 89.6%, followed closely by a Hampel estimator with 89.3%, the Bell estimator with 88.5%, and a bisquare ($k = 6.4$) with 87.5%.

[Table 6 about here]

Table 6, like Tables 4 and 5, reveals no uniformly superior estimator. However, given estimators with high overall performance, we may wish to examine the individual efficiencies for each estimator, particularly if we consider some sampling distributions to be more plausible than others. One reasonable approach is to sacrifice a percentage point or two of efficiency for the slash distribution in return for slightly better performance at the Gaussian and contaminated Gaussian distributions. Such an approach favors the Bell estimator and bisquare estimator with $k = 6.4$ over a Hampel estimator ($a = 1.7$) and bisquare estimator ($k = 6.0$), which have marginally higher asymptotic triefficiencies. Since the Bell estimator has a simpler definition and uniformly better small sample efficiency than the bisquare with $k = 6.4$ (which appears converge slowly to its asymptotic performance—see Holland and Welsch, 1977), the Bell estimator is a good overall choice. It should be borne in mind, however, that estimators with efficiencies differing by a percentage point or two may be regarded, for all practical purposes, as having identical efficiencies.[16]

Gross Error Sensitivity. A somewhat crude asymptotic measure of robustness is given by the *gross error sensitivity* γ^* (Hampel, 1974a). Ideally, one observation y added to a large sample should affect the estimator T_n only negligibly—an estimator T_n should summarize characteristics of the sample and not those of one particular observation. In the worst case, the estimated value for T_n is determined by one strategically placed observation. These two cases correspond to $\mathrm{IC}(y; F, T, S) = 0$ and $\mathrm{IC}(y; F, T, S) = \infty$, respectively. Hence, γ^* is defined as the largest absolute value attained by $\mathrm{IC}(y; F, T, S)$, that is,

$$\gamma^* = \sup_y |\mathrm{IC}(y; F, T, S)| . \tag{29}$$

The gross error sensitivity can also be used to give a rough approximation to the maximum bias of an estimator T. Suppose one observation is added to a large sample. Then the maximum bias of T is given asymptotically by $\epsilon\gamma^*$ since

$$\sup_y |T(F_\epsilon) - T(F)| \approx \epsilon\gamma^*$$

by the definition of the influence curve in (23).

Table 7 gives some values for γ^* for the estimators and distributions considered in Table 5. The mean has the worst attainable gross error sensitivity, with $\gamma^* = \infty$, since ψ

for the mean is unbounded. The median has the lowest values for γ^* among the estimators and the distributions listed in Table 7. Goodall (1983) notes that there is a rough trade-off between gross error sensitivity and efficiency and an examination of Tables 4–7 confirms that the good gross error sensitivity of the median is gained at the expense of lower overall efficiency. However, the estimators found to have high triefficiencies in Table 6 represent excellent compromises and have both high efficiency and low gross error sensitivity.

[Table 7 about here]

5. *HYPOTHESIS TESTING AND CONFIDENCE INTERVALS*

This section discusses tests of hypotheses and confidence intervals for M-estimators. While a better understanding of the underlying issues is beginning to emerge (see, for example, Efron, 1981; Iglewicz, 1983), this subject has received less attention than the robust estimation of location. Hence the methods discussed here must be viewed tentatively. In light of these difficulties, I have chosen to concentrate on relatively simple t-like tests similar to the classical Student t-test and a t-like test due to Johnson (1978; see also Efron, 1981).[17] These simple methods appear to perform well (Gross, 1976; Shorack, 1976; Efron, 1981; Martinez and Iglewicz, 1981; Kafadar, 1982; Shoemaker and Hettsmansperger, 1982; Iglewicz, 1983) and have the added advantage of familiarity and ease of usage.

The classical test of a hypothesis $H_0 : \theta = \theta_0$ versus an alternative hypothesis $H_1 : \theta > \theta_0$ involves the familiar Student t-test statistic

$$\frac{\overline{y} - \theta}{s_n / \sqrt{n}}, \tag{30}$$

where the estimators of location and spread are the sample mean $\overline{y}$ and the sample standard deviation s_n, respectively. Given a level α, the procedure typically used is to accept the alternative hypothesis H_1 if the statistic in (30) exceeds the critical value $\kappa_{\alpha,n-1}$ and otherwise to accept the null hypothesis H_0, where the critical value $\kappa_{\alpha,n-1}$ denotes the $(1-\alpha)100$ percentile of the Student t-distribution with $n-1$ degrees of freedom. Alternatively, one can use the statistic in (30) to form the classical $1-2\alpha$ central confidence interval

$$[\overline{y} - \kappa_{\alpha,n-1} s_n \, , \, \overline{y} + \kappa_{\alpha,n-1} s_n] \tag{31}$$

The value α is the level of the test and denotes the probability of a Type I error:

$$\Pr[\text{Reject } H_0 \mid H_0 \text{ true}] = \Pr[\sqrt{n}(\overline{y} - \theta)/s_n > \kappa_{\alpha,n-1}] = \alpha\,.$$

For Gaussian distributions, the test statistic in (30) can be shown to be the most powerful α-level test, that is, a test producing the shortest confidence intervals or, equivalently, minimizing the probability of a Type II error: $\Pr[\text{Accept } H_0 \mid H_0 \text{ false}]$.

How does the Student t-test perform if the underlying distribution differs from a Gaussian distribution? For customary levels of α, for example, $\alpha \leq .05$, the Student t-test typically yields a conservative test for most distributions, for example, distributions resembling a Gaussian distribution in the center but having greater mass in the tails (Benjamini, 1983). That is, the classical t-test may lead an analyst to reject the null hypothesis more often than would be expected from a .05 level test for heavy-tailed distributions. This is intuitively plausible because the standard deviation is highly nonrobust—a few aberrant observations cause s_n to become extremely large even if the bulk of the data conform to a Gaussian distribution. Since s_n enters into the denominator of t in (30), the value for t, and hence the level of the test, tends to be biased downward severely, leading to a conservative test. Moreover, the classical t-test appears to be even less robust with respect to the power of the test than to the level of the test (Hampel, 1973).

The test statistic in (30) suggests a robust test statistic

$$t = \frac{T_n - \theta}{A_n/\sqrt{n}}\,, \tag{32}$$

where T_n is the M-estimator and $A_n^2 = A^2(T_n; F_n, S_n)$ is the estimator of spread defined by a suitable modification of (28) and with the same M-estimator T_n. Under certain regularity conditions and for symmetric F, T_n and A_n^2 are independent and have asymptotic Gaussian and chi-square distributions, respectively (Huber, 1981). Then a standard theorem states that the distribution of the statistic in (32) should be close to a Student t-distribution with ν degrees of freedom (df). A key difficulty, however, concerns the appropriate value for ν, which appears to depend in a complicated fashion on the underlying distribution F, the shape of ψ, the value of the tuning constants, and the estimator of the auxiliary scale parameter S (Shoemaker and Hettsmansperger, 1982).

To date, two approaches have proven popular in constructing confidence intervals and t-like tests based on the statistic in (32). One approach (Gross 1976; Kafadar, 1982; Iglewicz, 1983) constructs special tables for the critical values of t using the results of Monte Carlo simulations for particular M-estimators and a few selected sampling distributions. Since the critical values of t vary for different distributions, a typical procedure is to take the largest critical value, which provides a conservative test. Iglewicz (1983) reports critical values for $\alpha = .05$ and samples sizes of 10, 20, 30, 40, 50, and 100 for the bisquare M-estimator with $k = 9.0$ and for sample sizes of 20 for the sine M-estimator with $k = 2.4\pi$. For both estimators, the conservative critical values occur for the Gaussian distribution.

A somewhat *ad hoc* but simpler solution followed in this chapter uses critical values $\kappa_{\alpha,\nu}$ from standard t-tables but reduces the classical $n - 1$ *df* by a constant fraction, where the specific fraction is usually determined by Monte Carlo simulation. For example, Mosteller and Tukey (1977) recommend $0.7(n-1)$ *df* for the bisquare. Martinez and Iglewicz (1981), Kafadar (1982), and Iglewicz (1983) find that using the test statistic in (32) with $0.7(n-1)$ *df* produces critical values that agree closely with those found in simulations for a bisquare estimator with $k = 9.0$. Martinez and Iglewicz (1981) suggest using $0.6(n-1)$ *df* for a sine M-estimator with $k = 3.1\pi$ and for a Hampel M-estimator with $a = 2.25, b = 3.75$, and $c = 15.0$.[18]

More formally, the tests considered above are based on a pair of estimators (T_n, A_n^2). As noted earlier, the center of the density was a natural location parameter under the assumption of a symmetric density f. But even under the assumption of a symmetric density f there is no such natural scale parameter, leading to some arbitrariness in the choice of the scale estimator A_n^2 (Iglewicz, 1983). One consequence is that the choice of the matching scale estimator, as in (32), does not guarantee a most powerful test in the sense of producing shortest confidence intervals or minimizing the probability of a Type II error. Despite these difficulties, evidence from numerous Monte Carlo studies (Gross, 1976; Shorack, 1976; Kafadar, 1981; Iglewicz, 1983) suggest that the test statistic in (32) performs well in practice in terms of both level and power.

Since a bisquare estimator with $k = 6.4$ and Bell estimator with $k = 1/0.35$ were found to perform better than the two estimators used by Kafadar and Iglewicz, I checked the performance of tests using $0.7(n-1)$ *df* for these estimators in simple Monte Carlo simulations. Following Rocke and Downs (1981) and Kafadar (1982), A_n^2 is defined to be

$$A_n^2 = \frac{(kS_n)^2 \sum \psi^2(u)}{[\sum \phi(u)][-1+\sum \phi(u)]}\,. \tag{33}$$

The denominator of (33) differs slightly from that suggested by the asymptotic expression in (28) in a manner similar to the small sample correction of $n-1$ in the definition of the sample variance.

Two test statistics were examined in the Monte Carlo simulations: the test statistic t in (32), and t^*, an analogue of a test statistic due to Johnson (1978; see also Efron, 1981), defined as

$$t^* = t + \frac{\mu_3}{3A_n^3\sqrt{n}}\left(t^2 + \frac{1}{2}\right) \tag{34}$$

where t is the test statistic in (32) and

$$\mu_3 = \frac{(kS_n)^3 \sum \psi^3(u)}{([\sum \phi(u)][-1+\sum \phi(u)])^{3/2}}\,. \tag{35}$$

Johnson uses a Cornish-Fisher expansion for the mean and standard deviation and finds that t^* has a distribution matching the Student t-distribution more closely than the classical t-statistic in (32). For the mean, μ_3 is the third moment of the distribution; (35) presents a similar expression for M-estimators suggested by the expression in (33) for A_n^2.[19]

Tables 8–10 report the observed levels for nominal 1%, 5%, 10%, and 15% two-sided tests using t and t^* for the bisquare and Bell estimators with $k = 6.4$ and $k = 1/0.35$, respectively. For comparison, results for the mean and classical Student t-statistic are also reported. I examined three sampling distributions: the standard Gaussian distribution, the 5% contaminated Gaussian at scale 10, and the slash distribution. Five thousand samples of size 20 were used in sampling from the Gaussian and 5% contaminated Gaussian distributions; 20,000 samples of size 20 were used in sampling from the slash distribution.

[Tables 8–10 about here]

As expected, the classical Student t-test for the mean performs well for the Gaussian distribution but produces a conservative test for the contaminated Gaussian and slash distributions. For a nominal 5% two-sided test, the observed levels for the Student t-test are approximately 3% and 2% for the contaminated and slash distributions, respectively.

The tests using $n - 1$ *df* also do not provide optimal results for the bisquare and Bell estimators. For example, the observed levels for the t-test in (32) for a nominal 5% two-sided test with 19 *df* are approximately 6% for the Gaussian and contaminated Gaussian distributions for both the bisquare and Bell estimators, which produces a somewhat liberal test. Since reducing the degrees of freedom from $n - 1$ to $0.7(n - 1)$ implies larger critical values, the resulting tests are more conservative and produce observed levels in Tables 8–10 that agree more closely with the nominal levels of α.[20] The two test statistics t and t^* differ only slightly for the Gaussian and contaminated Gaussian distributions when used with the bisquare and Bell M-estimators, with t^* tending to provide a slightly more conservative test than t. The differences are more marked for the slash distribution, where t^* tends to provide a less conservative test than t.

Since a conservative test is usually preferred to one that is liberal, the results in Tables 8–10 suggest that for the Bell estimator, using t^* with $0.7(n-1)$ degrees of freedom provides a good overall test that improves on the overly conservative Student t-test. However, a more systematic study is clearly required to determine the best degrees of freedom, evaluate the performance of different test statistics and estimators of spread, and assess the effects of asymmetric sampling distributions. In particular, t^* for the bisquare might perform better with a different adjustment to the degrees of freedom.

6. *ROBUST REGRESSION*

In this section, I consider M-estimation of the usual linear regression model[21]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is an $n \times p$ matrix of known values for p independent variables, $\mathbf{y}$ is an $n \times 1$ vector of observations for the dependent variable, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random disturbance terms ϵ_i, $i = 1, \ldots, n$. The ϵ_i are assumed to be i.i.d. with $\mathrm{E}(\epsilon_i) = 0$.

M-estimators of regression. The OLS estimator $\mathbf{b}_{\mathrm{LS}}$ is given by

$$\mathbf{b}_{\mathrm{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,.$$

(I denote any estimator of the unknown parameter vector $\boldsymbol{\beta}$ by $\mathbf{b}$.) This estimator is fully efficient if the ϵ_i conform to a Gaussian distribution. But like the mean, $\mathbf{b}_{\mathrm{LS}}$ lacks resistance, has a breakdown point of zero, and quickly loses efficiency for error distributions with heavier tails than the Gaussian distribution.

The usual linear model can be rewritten as

$$\epsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}, \qquad i = 1, \ldots, n\,,$$

where $\mathbf{x}_i$ is the ith row of the matrix $\mathbf{X}$. This suggests defining the regression M-estimator as the value of $\mathbf{b}$ that minimizes

$$\sum_{i=1}^{n} \rho(u_i)\,, \tag{36}$$

where

$$u_i = \frac{y_i - \mathbf{x}_i'\mathbf{b}}{kS}\,. \tag{37}$$

Taking partial derivatives of the expression in (36) with respect to b_j yields the p equations

$$\sum_{i=1}^{n} x_{ij}\,\psi(u_i) = 0 \qquad j = 1, \ldots, p\,, \tag{38}$$

where x_{ij} denotes the ijth entry of the matrix $\mathbf{X}$.

Rewriting the last expression in a form similar to the weighted mean in expression (18) gives

$$\sum_{i=1}^{n} x_{ij}\,u_i\,w_i = 0\,, \qquad j = 1, \ldots, p\,,$$

where the weights w_i are defined by

$$w_i = \psi(u_i)/u_i\,. \tag{39}$$

This yields a weighted least-squares expression (Holland and Welsch, 1977; Byrd and Pyne, 1979; Hogg, 1979)

$$\mathbf{b} = (\mathbf{X}'\,\mathbf{W}\,\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{W}\mathbf{y}\,, \tag{40}$$

where $\mathbf{W}$ is a diagonal matrix with diagonal elements w_i.

Estimation. As in the case of estimating a location parameter T_n, one can either minimize the expression in (36) using a Newton–Raphson type algorithm or apply the iteratively reweighted least-squares (IRLS) method suggested by (40) to obtain numerical values for the M-estimator $\mathbf{b}$. Newton–Raphson methods typically employ a search procedure to identify the steepest descent direction at each iteration; the Hessian matrix may be used to determine if convergence is to a local minimum or saddle point of $\sum \rho(u)$. For the IRLS procedure, the initial estimate $\hat{\mathbf{b}}^{(0)}$ is used to calculate the weights w_i, which are in turn used to calculate a new value $\hat{\mathbf{b}}^{(1)}$, continuing until a convergence criterion is satisfied. Byrd and Pyne (1979) discuss conditions under which the IRLS method converges either to a saddle point or local minimum. They show that under mild conditions the IRLS method does not converge to a local maxima in general. Convergence proofs generally assume that the auxiliary parameter of scale S_n is either known or fixed; however, one also typically iterates on the auxiliary scale parameter, particularly in the case of estimating regression parameters.

Hogg (1979) notes that robust estimation of the nonlinear model $\mathbf{y} = g(\mathbf{x}_i, \mathbf{b}) + \epsilon$ follows from a simple modification of the u_i

$$u_i = \frac{y_i - g(\mathbf{x}_i, \mathbf{b})}{kS},$$

where $g(\mathbf{x}_i, \mathbf{b})$ is a nonlinear function. Then M-estimates are obtained by either minimizing $\sum \rho(u_i)$ or using an IRLS method obtained by taking partial derivatives of $\sum \rho(u_i)$ to calculate the appropriate weights for the expression in (40).

Leverage Points and Multiple Solutions. The data in Figure 1 considered at the outset of this chapter illustrated an example of the problems caused by leverage points, that is, outliers in the x's that can (potentially) exert a strong influence on the parameter estimates $\hat{\mathbf{b}}$ or on the predicted values $\hat{y}_i$ by virtue of their position in the data. Both the OLS estimator and the Bell estimator started from the OLS solution were adversely affected by the outlying cluster of leverage points. However, the Bell estimator started from the repeated median estimate identified and downweighted the cluster of 20 "bad" points.

The results in Figure 1 showed that the performance of the regression M-estimator depends heavily on the breakdown properties of the initial estimator. Consider the Bell estimator started from the OLS estimate. Such a start provided 50 moderate-sized initial estimates of the u_i. In this situation, the Bell estimator could not identify the 20 outlying observations and hence yielded estimates similar to the OLS estimates. In contrast, the repeated median estimate provided the Bell estimator with 30 small values of u_i and 20 large values of u_i. This allowed the Bell estimator to refine the rough initial estimate ($\hat{a} = 3.28$, $\hat{b} = 0.816$) by downweighting the 20 leverage points, thus yielding a more precise estimate ($\hat{a} = 2.04$, $\hat{b} = 0.999$).

The expression in (38) helps to explain why the Bell estimator started from the repeated median estimate performed well. For redescending M-estimators like the Bell M-estimator, the function $\psi(u)$ rapidly approaches zero as $\pm u \to \infty$. Since an extreme outlier in the x's produces a large value for u_i, the resulting value of $\psi(u_i)$ is effectively zero. Hence a large value of x is compensated by a nearly zero value of $\psi(u)$ when minimizing the expression $\sum x\psi(u)$ in (38). Then, given a sufficiently high breakdown initial estimate, a redescending ψ function allows the regression M-estimator to downweight leverage points and other extreme outliers and to have the same high breakdown as the corresponding location M-estimator (Donoho, 1984).[22]

Poor breakdown may also explain the performance of the diagnostic procedures for the data in Figure 1. Gasko and Donoho (1982) find that many such procedures have surprisingly low breakdown points. Hence, these diagnostic procedures can fail to identify outliers when the data contain even small clusters of leverage points. This suggests that the breakdown properties of regression estimators or diagnostic procedures yield important information about their performance when the data contain severe contamination such as clusters of leverage points.

As the example in Figure 1 also demonstrates, one difficulty with redescending M-estimators is the possibility of multiple solutions. As noted earlier, although a single estimate for a given set of data and model is attractive, an implicit assumption is that the data in fact provide a single unitary indication. But multiple solutions often arise if the error distribution contains multiple modes. Hence, the choice between different estimates involves issues that cannot be easily resolved by simple statistical criteria. For example, the M-estimate started from the OLS estimate had a *lower* value of $\sum \rho(u_i)$ than the M-estimate started from the repeated median estimate, even though the latter possessed a higher breakdown point and so summarized the majority of the data better. Thus, automatically accepting the estimate that minimizes $\sum \rho(u)$ may not provide a reliable guide for choosing between different estimates. Instead, multiple solutions often indicate the need for model criticism or a reevaluation of substantive theory. For example, the goal of fitting a straight line to all observations in Figure 1 is clearly inappropriate since the data consist of two extremely disparate populations, a fact not reflected in the simple linear model used in the example.

It should be stressed that the M-estimators of regression defined in (36) and (38) are designed to be robust against deviations from the assumed distribution of the ϵ_i's. Thus, an implicit assumption is that the matrix $\mathbf{X}$ is known and observed without error. Notwithstanding the conceptual distinction between outliers in the y's and outliers in the x's (including leverage points), it is often difficult in practice to distinguish between the two since both produce large values for the u_i in (37) given a sufficiently robust (high breakdown) initial estimator.

Testing. Simple tests of hypotheses about individual parameters b_j of the form $H_0\colon b_j = \beta_j$ versus an alternative hypothesis $H_1\colon b_j > \beta_j$ generalize in a straightforward manner from the tests associated with a location estimator (Hogg, 1979; Huber, 1981). The classical test forms the t-statistic

$$t_{b_i} = \frac{b_j - \beta_j}{s_{b_j}}, \tag{40}$$

where $s^2_{b_i}$ is defined by the jth diagonal element of the covariance matrix

$$\mathrm{Cov}(\mathbf{b}_{\mathrm{LS}}) = \frac{s^2_n}{n-p}(\mathbf{X}'\mathbf{X})^{-1},$$

and s^2_n is defined by

$$s^2_n = \sum_{i=1}^{n}(Y_i - \mathbf{x}'_i\,\mathbf{b})^2 .$$

An analogous robust test for $\mathbf{b}$ replaces s_{b_j} in (41) by the square root of the jth diagonal of the covariance matrix

$$\mathrm{Cov}(\mathbf{b}) = \frac{A^2_n}{n-p}(\mathbf{X}'\mathbf{X})^{-1},$$

where A^2_n is defined in (33), and u_i is given in (37). The resulting value of t can then be used to evaluate the modification of Johnson's t-statistic given in expression (34).

7. *EXAMPLES*

The section presents two empirical examples illustrating the robust regression M-estimators described in the previous section. The first example consists of a simple regression using data on aggregate employment in metropolitan areas previously analyzed in a study of urban growth (Norton, 1979). The second example uses data from the 1979 wave of the National Longitudinal Survey of Youth (Borus and Santos, 1983) and examines individual earnings of young white males employed and out of school at the date of survey.

Growth of Employment in the Service Sector. The urban data describe growth of employment in the service sectors of the 30 largest Standard Metropolitan Statistical Areas for the period 1947–1972. These data provide an interesting example since the cities sampled represent a heterogeneous population consisting of older industrial cities (for example, Boston, Cleveland, Pittsburgh) and younger and rapidly growing sun-belt cities (for example, Dallas, Houston, Phoenix). In addition, Norton identifies an obvious outlier (Phoenix) and presents two sets of parameter estimates—OLS estimates for the full sample and OLS estimates after deleting data for Phoenix. His procedures provide a convenient comparison with robust methods, which are designed to accommodate and identify outliers in an automatic fashion.

Table 11 presents results for the specification used by Norton. The percentage growth of employment in the service sector is regressed on the percentage growth of employment in the manufacturing sector (for details on the definition of these variables, see Norton, 1979, p. 109). The OLS results for the full sample ($n = 30$) indicate that cities experiencing no growth in manufacturing employment nevertheless experienced an average rate of growth of 61 percent in service sector employment for the period 1947–1972. Similarly, a one percent increase in the growth of manufacturing employment resulted in an increase of 0.4% in the growth of service sector employment. All parameter estimates reported in Table 11 differ significantly from zero at the .05 level.

[Table 11 about here]

For the full ($n = 30$) sample, the Bell M-estimates started from the repeated median estimate (Bell/RM) differ considerably from the OLS estimates. Πhe Bell/RM estimate of the intercept implies an increase of 48% in service sector employment for cities experiencing no growth in manufacturing and differs from the OLS estimate by about two OLS standard errors. The Bell/RM estimate for sectoral employment implies that a one percent increase in manufacturing employment corresponds to an increase of 0.6% in the average growth rate of service sector employment, a value substantially larger than the OLS estimate.

Figure 8 presents a scatterplot of the 30 data points with the regression lines given by the OLS and Bell/RM regression estimates superimposed over the data. Of particular interest is the high leverage position of Phoenix, which experienced an extremely rapid rate of industrial growth during this period. The scatterplot clearly shows the leverage effect of the observation for Phoenix on the OLS estimate as well as the insensitivity of the robust M-estimator to this and other outlying observations.

[Figure 8 about here]

Figure 9 presents a Gaussian probability plot for the residuals obtained from the OLS and Bell/RM estimators. Such plots are often used to assess the fit between the assumed and sample distributions of the residuals (see, for example, Cook and Weisberg, 1982b, pp. 55–58). Residuals are plotted along the y-axis and the expected Gaussian quantiles along the x–axis. Note that the pattern of the robust residuals is close to linear, suggesting that the bulk of the residuals conform to a Gaussian distribution. (The robust residual for Phoenix is not plotted as it extends beyond the limits of the graph.) In contrast, the pattern of the OLS residuals deviates markedly from linearity. For both estimates, the pattern of residuals indicates a small number of outliers since the residuals have an inverted "S" shape pattern in which large negative values of the residuals fall below and large positive residuals lie above the bulk of residuals.

[Figure 9 about here]

Because of the outlying position of Phoenix, Norton dropped it and obtained OLS estimates for the remaining sample ($n = 29$). For this sample, the OLS estimate of the intercept (see Table 11) agrees closely with the Bell/RM estimate, but the OLS estimate of the effect of change in manufacturing employment is slightly higher than the Bell/RM estimate. Since the Bell/RM estimator assigned Phoenix a weight close to zero ($\tilde{w}_i =$ .00041) in the full sample, the Bell/RM estimates for the sample with Phoenix deleted are virtually identical to the Bell/RM estimates for the full sample.

Table 12 lists cities with large residuals for the OLS and Bell/RM estimates for the $n = 30$ sample. Surprisingly, the magnitudes of the OLS residuals for three cities (San Diego, Houston, and Atlanta) are larger than the magnitude of the OLS residual for Phoenix. Note also that Dallas is not identified as a possible outlier by the OLS estimates in the full robust standard error). Because of this, the effects for collective bargaining and urban residence achieve significance at the .05 level in the Bell estimates but only approach significance for the OLS estimates ($p < .15$). Thus, despite overall similarities, the differences in significance levels could lead to quite different substantive conclusions.

Are these lower levels of significance due to the smaller magnitudes of the OLS estimates or the smaller standard errors of the Bell estimates? A rough indication is provided by computing the ratios of the different parameter and standard error estimates for collective bargaining and urban residence. For both variables, the ratio of the OLS estimate to the Bell standard error is less than 1.7 while the ratio of the Bell estimate to the OLS standard errors is slightly greater than 2.0. This suggests that the differences between the magnitudes of the OLS and Bell coefficients are large enough to account for the differences in significance levels.

In order to investigate the shifts in the coefficient estimates, I examined partial regression plots (Mosteller and Tukey, 1977; Belsley, Kuh, and Welsch, 1980; Henderson and Velleman, 1981) for collective bargaining and urban residence. Given a model $y = b_0 + b_1 x_1 + \cdots + b_p x_p + e$, let $e(y_{\cdot 023 \ldots p})$ be the residuals obtained by regressing y on all independent variables except x_1 and let $e(x_{1\ 023 \ldots p})$ be the residuals obtained from regressing x_1 on $x_2, \ldots, x_p$. Then the partial regression plot for b_1 is obtained by plotting $e(y_{\cdot 023 \ldots p})$ against $e(x_{1\ 023 \ldots p})$. Where no confusion may occur, I denote $e(x_{1\ 023 \ldots p})$ and $e(y_{\ 023 \ldots p})$ by $e(x_{1\ \text{rest}})$ and $e(y_{\cdot \text{rest}})$, respectively. Mosteller and Tukey (1977) show that slope of the OLS line through the points of the partial regression plot equals the OLS estimate of b_1. Moreover, the OLS residuals of the partial regression plot are the same as the residuals in the original model. Thus, these plots provide an extremely useful tool by allowing one to inspect the effects of individual observations on a particular coefficient.

Figure 11 illustrates the partial regression plot for collective bargaining using the restricted ($n = 119$) sample. The solid line is the OLS line fitted to the points in the partial regression plot; the slope of this line equals the OLS estimate for the effect of collective bargaining in the $n = 119$ sample. For comparison, I have superimposed a dashed line whose slope equals the robust parameter estimate; note, however, that this line need not coincide with the M-estimate fitted to these points, as is the case for the OLS estimate.

[Figure 11 about here]

Visual inspection suggests a few unusual points are potential leverage points for this parameter. These points are labeled by an identifier and the weight assigned by the robust estimator. Considered separately, observations 14 and 99 are potential leverage points since the extreme positions of each observation along the x- and y-axes could cause the OLS estimate to shift. However their effects appear to cancel. Similarly, observation 44 has a large y-deviation but a more central x-position and so seems less influential. More problematic are the three observations located in the lower right-hand portion of Figure 11. Because these observations have relatively large residuals, the Bell estimator downweights the influence of these observations on the Bell estimate. But because the OLS estimator assigns equal weights to all observations, these observations appear to pull the estimated OLS line downward as compared to the robust line, which may explain the lower level of significance for the OLS estimate.

The original data show that the individuals in question reported lower than average hourly wages ($1.79, $2.90, and $2.56 for observations 56, 107, and 112, respectively). Althogh these wages seem low for jobs covered by collective bargaining, it is difficult to assert confidently that the observed data are in error and should be rejected. The Bell estimator provides a middle ground between complete acceptance and complete rejection of these observations by downweighting their influence, thus limiting but not eliminating the information in them.

The partial regression plot in Figure 12 for urban residence shows a similar overall pattern. While several large residuals occur in Figure 12, the positions of three observations (44, 56, and 99) appear to exert damaging effects on the OLS parameter estimate, resulting in a smaller estimated slope for the effect of urban residence. The small weights assigned to these observations, (.299 for observation 44; .320 for observation 56; .295 for observation 99) indicate that robust estimator provides a greater degree of protection against the possible downward bias caused by the strategic position of these observations than the OLS estimator. Once again, the three observations identified as potentially troublesome report unusually low hourly wages but, as before, it is difficult to assert with great confidence that such values are definitely inconsistent with the reported data for urban residence.

[Figure 12 about here]

Lastly, Table 14 provides a crude test of the breakdown properties of the OLS and Bell estimators by presenting results for the raw (untransformed) wage data using the full ($n = 120$) sample. Since these data have a pronounced positive skew and include the extreme outlier, this example presents a situation that departs sharply from the usual distributional assumptions.

Not surprisingly, the OLS estimate breaks down dramatically, yielding extremely aberrant results. All coefficients show large shifts from the results in Table 13 and only the effect of marital status remains significant at the .05 level. However, the Bell estimates, standard errors, and significance levels are strikingly similar to those in Table 13. The Bell estimates differ most for collective bargaining and marital status, with values roughly 0.75–1.0 standard errors higher in Table 14 than in Table 13. The larger coefficients for collective bargaining and marital status are not surprising in light of the positive skewness of the distribution of raw wages.

[Table 14 about here]

Obviously, only a naive analyst would fail to omit the obvious outlier and transform the dependent variable to achieve symmetry, particularly since these features are readily apparent in even a cursory examination of the stem and leaf display in Figure 10. Nevertheless, it is reassuring that the robust M-estimator yields reasonable estimates even in the presence of such noticeable departures from the distributional assumptions.

8. *CONCLUSION*

Careful data analysts have long been alert to the pitfalls resulting from unusual observations and distributions with greater mass in the tails than a Gaussian distribution. Unfortunately, the classical estimators routinely used by sociologists perform poorly for distributions that deviate even slightly from a Gaussian distribution and may produce parameter estimates that are misleading or useless. This is of considerable practical concern since empirical data seldom—if ever—conform *exactly* to a Gaussian distribution.

This chapter has discussed and illustrated one class of robust estimators, the M-estimators, which are designed to perform well for a wide range of distributions. The considerations in Section 4 showed that M-estimators satisfy several common statistical criteria, are resistant to large changes in a few observations or small changes in many observations, have a high breakdown point against severe departures from the distributional assumptions or large contaminations of the data, and perform efficiently for distributions resembling a Gaussian distribution in the center but differing in the tails. Moreover, M-estimators are readily extended to problems such as multiparameter regression and tests of hypotheses.

Because extreme outliers are easily identified by a variety of methods and can typically be rejected, they seldom cause great difficulties. Hence, alternatives to M-estimators, such as regression diagnostic procedures and graphical displays (scatterplots, stem and leaf displays, partial regression plots), are often useful for identifying extreme outliers or other unusual observations. Still, care must be taken. The usual practice of inspecting the OLS residuals can fail to identify outliers, as the urban data and the example in Figure 1 illustrated, even when the outliers are easily identified in graphical displays. Regression diagnostic procedures can also fail to identify extreme outliers in certain extreme situations, as the example in Figure 1 illustrates. Graphical displays are typically revealing but may occasionally be misleading in complicated problems involving many independent variables if the outliers do not have extreme values on any one variable (Friedman and Stuetzle, 1982).

More problematic are situations in which one or more observations exert moderate but potentially damaging effects on the parameter estimates, standard errors, or significance levels. The NLS data on wages provide two informative examples in which the OLS and robust M-estimates differ in significance level. A more careful analysis of these data suggests that a few individuals with low reported wages appear to exert a downward pull on the OLS estimates but not on the M-estimates. Although the estimates are roughly similar in magnitude, the differences are great enough to account for the lower levels of significance given by the OLS estimator as compared to the M-estimator ($p < .15$ vs. $p < .05$). Clearly, these differences could easily alter the conclusions drawn from these data on two variables of considerable substantive interest.

Although regression diagnostic procedures and graphical displays can often identify possibly influential observations in such intermediate situations, the problem of what to do with such observations once identified is less straightforward. One approach is to sequentially delete selected observations or subsets of observations and observe the effects on the parameter estimates. But such a process is arduous, particularly if many observations are singled out for deletion or if (as in the NLS data) the observations identified as influential vary across parameters. Moreover, we rarely have definitive substantive guidance for rejecting or accepting borderline cases, which can render difficult data-cleaning decisions even more difficult.

Because the M-estimators provide a formal procedure for progressively downweighting observations, they allow one to avoid the all or nothing approach of methods that clean the data. Of course, the additional effort needed to obtain the robust M-estimators may seem superfluous if the conclusions are unaffected or change in only small ways. But the examples indicate that the results obtained from the robust estimators can have quite different and important substantive implications.

Despite many advantages, robust estimators have certain drawbacks. They must be estimated by an iterative procedure and hence are more costly to compute than the classical estimators. The use of a robust initial estimate is highly desirable, particularly when multiple solutions exist, but is costly to obtain in large samples with many independent variables. Although the robust tests proposed in Sections 5 appeared to work well for sample sizes of $n \geq 20$, more work is needed to develop better and more efficient methods for testing hypotheses and constructing confidence intervals. Finally, although the IRLS estimation procedure lets one adapt existing OLS procedures to M-estimators of regression, the present lack of readily available software poses obstacles to widespread use.

Should we abandon classical estimators like the sample mean and OLS estimators? Although the answer is a qualified no, we would do well to modify our usual practices. One reasonable approach is to examine the results obtained from both the classical and robust estimators. If the results agree substantially, we should report the agreement and the results of either (or both) procedures. If the results disagree, further analysis can often help to isolate the causes of the differences, and the residuals or weights produced from the robust procedure can help to identify unusual observations. Such observations can be checked for error or influence on the estimated parameters or predicted values. They also often help to suggest deficiencies in the model.

Based on a variety of theoretical and practical criteria, the redescending M-estimators are particularly useful. Because of excellent overall performance (resistance, high breakdown, high triefficiency, and low overall gross error sensitivity), the Bell M-estimator represents a good choice among the more popular redescending M-estimators. However, the differences between the better redescending M-estimators are often slight and other choices, such as Tukey's bisquare with tuning constant 6.0 or 6.4, are likely to give results similar to those obtained with the Bell M-estimator. It is less easy to make definitive recommendations regarding procedures for constructing confidence intervals and tests of hypotheses, partly because these topics are less well understood and much more work remains to be done. However, the results of preliminary Monte Carlo simulations reported in Section 4 indicate that tests using a modification of Johnson's t-statistic with $0.7(n-1)$ df perform well when coupled with the Bell M-estimator.

Strong negative recommendations are made most easily. The worst estimators and tests—by substantial margins—are the classical estimators and tests. If the data conform exactly to the classical assumptions, then one loses little by using robust estimators and tests. But one can lose substantially by only using the classical methods if the data depart even slightly from the classical distributional assumptions.

APPENDIX 1

This appendix presents definitions for the least median of squares (Rousseeuw, 1982) and repeated median (Siegel, 1982) estimators for **b** in the linear model

$$y_i = \mathbf{x}_i'\mathbf{b} + \epsilon_i \, .$$

Although neither estimator has particularly high efficiency for Gaussian error distributions, either can be used to provide a highly robust (high breakdown) initial estimate. Both estimators are costly to compute for large n and many independent variables.

The least median of squares estimator is defined by the value of **b** minimizing

$$\operatorname*{median}_{i} \, (y_i - \mathbf{x}_i^T \mathbf{b})^2 \, , \tag{A1}$$

where $i = 1, \ldots, n$. In the case of the simple regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the minimization problem in (A1) corresponds to finding the narrowest strip that covers half of the observations (Rousseeuw, 1982). Somewhat unexpectedly, the least median of squares estimator has a slower rate of convergence than (the usual) $n^{1/2}$ and converges to its asymptotic performance at the rate $n^{1/3}$.

The repeated median estimator (Siegel, 1982) is defined by a series of nested medians. In the case of the simple regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the repeated median estimate for β_1 is defined by

$$\beta_1 = \operatorname*{median}_{i_1} \; \operatorname*{median}_{i_2 \neq i_1} \; \frac{y_{i_2} - y_{i_1}}{x_{i_2} - x_{i_1}} \, , \tag{A2}$$

and the repeated median estimate for β_0 by

$$\beta_0 = \operatorname*{median}_{i_1} \; \operatorname*{median}_{i_2 \neq i_1} \; \frac{x_{i_2} y_{i_1} - x_{i_1} y_{i_2}}{x_{i_2} - x_{i_1}} \, , \tag{A3}$$

where i_1 and i_2 range over $i = 1, \ldots, n$.

More generally, let **X** be an $n \times p$ matrix for p independent variables and let the indices $i_1, \ldots, i_p$ range over $i = 1, \ldots, n$. Then the repeated median estimate for **b** is given by

$$\mathbf{b} = \operatorname*{median}_{i_1} \; \operatorname*{median}_{i_2 \notin \{i_1\}} \cdots \operatorname*{median}_{i_p \notin \{i_1, \ldots, i_{p-1}\}} \; \mathbf{b}(i_1, \ldots, i_p) \, . \tag{A4}$$

The vector $\mathbf{b}(i_1, \ldots, i_p)$ denotes the (unique) solution of the system of p equations with p unknowns

$$\begin{aligned} y_{i_1} &= b_1 x_{i_1,1} + b_2 x_{i_1,2} + \cdots + b_p x_{i_1,p} \\ &\vdots \qquad\qquad\qquad \vdots \\ y_{i_p} &= b_1 x_{i_p,1} + b_2 x_{i_p,2} + \cdots + b_p x_{i_p,p} \,, \end{aligned}$$

where x_{ij} is the ij-th entry of the matrix $\mathbf{X}$. The innermost median of (A4) lets the index i_p range over all values of $i = 1, \ldots, n$ not equal to $i_1, \ldots, i_{p-1}$ and takes the element-wise median of the $n - p + 1$ values of $\mathbf{b}(i_1, \ldots, i_{p-1}, \cdot\,)$. Linear dependencies in the p equations can be handled by considering only those p-tuples of observations that determine a unique value for $\mathbf{b}(i_1, \ldots, i_p)$.

Both estimators can be shown to have breakdown points that approaches 50% as $n \rightarrow \infty$ for finite p. While both estimators are based on medians, not all median-based regression methods have high breakdown. For example, least absolute value regression has a 50% breakdown point for outliers in the y's but 0% breakdown for outliers in the x's.

The computational complexity of the repeated median estimator requires $O(n^p)$ operations; note that the median can be found in $O(n)$ operations. I suspect that the computational complexity of the least median of squares estimator is at least as great as that for the repeated median estimator.

APPENDIX 2

In this appendix, I give an informal derivation of the influence curve for M-estimators; for rigorous proof, see Huber (1981). Recall the definition of the influence curve in expression (23)

$$\mathrm{IC}(y; F, T, S) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon(\nu)) - T(F(\nu))}{\epsilon} \,, \tag{A5}$$

where

$$F_\epsilon(\nu) = (1 - \epsilon) F(\nu) + \epsilon \xi_y(\nu) \,, \tag{A6}$$

$F(\nu)$ is the underlying distribution, ϵ is a small positive number, and ξ is the indicator function defined in expression (5). As before, I have replaced the usual $F(y)$ by $F(\nu)$ for notational clarity. Note that $F(\nu) = [F_\epsilon(\nu)]_{\epsilon=0}$. I wish to show that

$$\mathrm{IC}(y; F, T, S) = \frac{kS(F)\,\psi\Big(\dfrac{y - T(F)}{kS(F)}\Big)}{\displaystyle\int \phi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\,\mathrm{d}F(\nu)}\,, \qquad (A7)$$

where F is symmetric and ψ is an odd function.

Two intermediate steps are needed to prove (A7). I first verify the following identity for $F_{h+\epsilon}(\nu)$

$$F_{h+\epsilon}(\nu) = \Big(1 - \frac{\epsilon}{1-h}\Big)F_h(\nu) + \frac{\epsilon}{1-h}\,\xi_y(\nu) \qquad (A8)$$

where h is a small positive number. Substituting (A6) into (A8) yields

$$\begin{aligned}
F_{h+\epsilon}(\nu) &= (1 - (h+\epsilon))F(\nu) + (h+\epsilon)\xi_y(\nu) \\
&= (1 - (h+\epsilon))F(\nu) + \frac{1}{1-h}(h(1-h) + \epsilon(1-h))\xi_y(\nu) \\
&= (1-h)F(\nu) - \epsilon F(\nu) + h\xi_y(\nu) + \frac{\epsilon}{1-h}\,\xi_y(\nu) - \frac{\epsilon h}{1-h}\,\xi_y(\nu) \\
&= \Big(1 - \frac{\epsilon}{1-h}\Big)\Big[(1-h)F(\nu) + h\xi_y(\nu)\Big] + \frac{\epsilon}{1-h}\,\xi_y(\nu) \\
&= \Big(1 - \frac{\epsilon}{1-h}\Big)F_h(\nu) + \frac{\epsilon}{1-h}\,\xi_y(\nu)\,.
\end{aligned}$$

This verifies (A8).

Second, I show that

$$\mathrm{IC}(y; F, T, S) = \Big[\frac{\partial T(F_h)}{\partial h}\Big]_{h=0}\,. \qquad (A9)$$

By the definition of a derivative for an implicit function

$$\Big[\frac{\partial T(F_h)}{\partial h}\Big]_{h=0} = \Big[\lim_{\epsilon\to 0}\frac{T(F_{h+\epsilon}) - T(F_h)}{\epsilon}\Big]_{h=0}\,.$$

To verify (A9), substitute the expression for $F_{h+\epsilon}$ in (A8) into the left-hand side of the above

$$\begin{aligned}
\Big[\frac{\partial T(F_h)}{\partial h}\Big]_{h=0} &= \Big[\lim_{\epsilon\to 0}\frac{T[(1 - \epsilon/(1-h))F_h + (1 - \epsilon/(1-h))\xi_y(\nu)] - T(F_h)}{\epsilon}\Big]_{h=0} \\
&= \lim_{\epsilon\to 0}\frac{T[(1-\epsilon)F_{h=0} + \epsilon\xi_y] - T(F_{h=0})}{\epsilon} \\
&= \lim_{\epsilon\to 0}\frac{T(F_\epsilon) - T(F)}{\epsilon} \\
&= \mathrm{IC}(y; F, T, S)
\end{aligned}$$

by the definition of the influence curve in (A5). Note, however, that some risky and possibly illegal interchanges of passages to the limit and evaluating the expression at $h = 0$ are involved in the above.

I next calculate the influence function for an M-estimator. If ψ is an odd function and the distribution F is symmetric about θ, then asymptotically an M-estimator $T(F)$ satisfies

$$\int \psi(u)\, \mathrm{d}F(\nu) = 0\,,$$

where

$$u = \frac{\nu - T(F)}{kS(F)}\,. \tag{A10}$$

Under certain mild regularity conditions (see Huber, 1981), one can substitute $F_h(\nu)$ for F into the above and differentiate the result with respect to h

$$\frac{\partial}{\partial h}\int \psi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)\, \mathrm{d}F_h(\nu) = 0\,. \tag{A11}$$

By the chain rule

$$\frac{\partial\psi(u)}{\partial h} = \frac{\partial\psi(u)}{\partial u}\frac{\partial u}{\partial h} = \phi(u)\frac{\partial u}{\partial h}\,.$$

Then replacing $T(F)$ by $T(F_h)$ in the expression for u in (A10) above yields

$$\frac{\partial}{\partial h}\psi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big) = \phi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)\frac{\partial}{\partial h}\Big[\frac{\nu - T(F_h)}{kS(F_h)}\Big]\,.$$

But

$$\frac{\partial}{\partial h}\Big[\frac{\nu - T(F_h)}{kS(F_h)}\Big] = -\frac{1}{kS}\frac{\partial T(F_h)}{\partial h} + \frac{\nu - T(F_h)}{kS^2(F_h)}\frac{\partial S(F_h)}{\partial h}\,.$$

Thus

$$\frac{\partial}{\partial h}\psi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big) = -\frac{1}{kS(F_h)}\phi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)\frac{\partial T(F_h)}{\partial h} + \frac{\nu - T(F_h)}{kS^2(F_h)}\phi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)\frac{\partial S(F_h)}{\partial h} \tag{A12}$$

Note also that

$$\frac{\partial}{\partial h}\mathrm{d}F_h(\nu) = \frac{\partial}{\partial h}\mathrm{d}[(1-h)F(\nu) + h\xi_y(\nu)] = \mathrm{d}[\xi_y(\nu) - F(\nu)] = (\delta_y(\nu) - f(\nu))\mathrm{d}\nu\,, \tag{A13}$$

where $\delta_y(\nu)$, denoting the Dirac delta function, is the derivative of the indicator function $\xi(\nu)$. Taking the partial derivative inside the integral (A11) gives

$$0 = \int \frac{\partial}{\partial h}\Big[\psi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)dF_h(\nu)\Big]\,.$$

Then substituting (A12) and (A13) into the above yields

$$\begin{aligned} 0 = &- \frac{1}{kS(F_h)} \frac{\partial T(F_h)}{\partial h} \int \phi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big) \mathrm{d}F_h(\nu) \\ &+ \frac{1}{kS^2(F_h)} \frac{\partial S(F_h)}{\partial h} \int (\nu - T(F_h))\phi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big) \mathrm{d}F_h(\nu) \qquad (A14) \\ &+ \int \psi\Big(\frac{\nu - T(F_h)}{kS(F_h)}\Big)(\delta_y(\nu) - f(\nu))\mathrm{d}\nu\,. \end{aligned}$$

For an even function ϕ and symmetric distribution F, $T(F_n)$ and $S(F_n)$ are asymptotically independent, and so the second integral in A(14) is identically zero. Then eliminating the second integral and letting $h \to 0$ yields

$$0 = -\frac{1}{kS(F)}\Big[\frac{\partial T(F_h)}{\partial h}\Big]_{h=0} \int \phi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\mathrm{d}F(\nu) + \int \psi\Big(\frac{\nu - T(F)}{kS(F)}\Big)(\delta_y(\nu) - f(\nu))\mathrm{d}\nu\,. \qquad (A15)$$

The second integral in (A15) can be expanded

$$\int \psi\Big(\frac{\nu - T(F)}{kS(F)}\Big)(\delta_y(\nu) - f(\nu))\mathrm{d}\nu = \int \psi(u)\delta_y(\nu)\mathrm{d}\nu - \int \psi(u)\mathrm{d}F(\nu)\,. \qquad (A16)$$

But by the definition of an M-estimator, the integral of $\psi(u)$ over $\mathrm{d}F(\nu)$ is identically zero. Hence, the second term on the right-hand side vanishes and (A16) simplifies

$$\int \psi(u)(\delta_y(\nu) - f(\nu))\mathrm{d}\nu = \int \psi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\delta_y(\nu)\mathrm{d}\nu\,. \qquad (A17)$$

Then since $\delta_y(\nu)$ is a delta function, the integral in (A17) is simply the value of ψ evaluated at $\nu = y$

$$\int \psi(u)\delta_y(\nu)\mathrm{d}\nu = \psi\Big(\frac{y - T(F)}{kS(F)}\Big)\,. \qquad (A18)$$

Hence, (A15) reduces to

$$0 = -\frac{1}{kS(F)}\Big[\frac{\partial T(F_h)}{\partial h}\Big]_{h=0} \int \phi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\mathrm{d}F_h(\nu) + \psi\Big(\frac{y - T(F)}{kS(F)}\Big)\,.$$

Then rearranging terms and using the identity in (A9) gives the desired result in (A7)

$$\mathrm{IC}(y; F, T, S) = \frac{kS(F)\,\psi\Big(\frac{y - T(F)}{kS(F)}\Big)}{\int \phi\Big(\frac{\nu - T(F)}{kS(F)}\Big)\,\mathrm{d}F(\nu)}\,.$$

APPENDIX 3

This appendix reproduces a theorem and proof due to Huber (1983b) regarding the breakdown point of redescending M-estimators. While the results are important and informative, they are as yet unpublished and so somewhat inaccessible.

The theorem shows that the breakdown point of a redescending M-estimator depends on the shape of the function ρ (or ψ), the value of the tuning constant k, and the sample Y, under the assumption that the auxiliary scale parameter S is fixed. Huber states that if the tuning constant is chosen such that efficiency for the Gaussian distribution is reasonably high and the gross error sensitivity is low, then the breakdown point of redescending M-estimators is quite high ($\epsilon^* > .40$ in most cases). If the estimator of scale is the MAD, then according to unreported numerical examples, Huber finds that the breakdown point can be close to optimal, for example, $\epsilon^* = .49$ for the bisquare with $k = 6$.

Let $Y = (y_1, \ldots, y_n)$ be a fixed sample of size n and let $Z = (z_1, \ldots, z_m)$ consist of m arbitrary values. I consider samples $Y^* = Y \cup Z$, that is, samples corrupted by ϵ-contamination, where $\epsilon = m/(n+m)$ is the proportion of contaminated points. For the purposes of the proof to follow, it is convenient to assume that (1) $\rho(u)$ is a monotone increasing even function with minimum $\rho(u) = \rho(0) = -1$ and (2) $\lim_{u\to\infty} \rho(u) = 0$. (Recall that ρ is defined only up to arbitrary additive and multiplicative constants.) In a slight abuse of notation, let $T(Y)$ denote the M-estimator for the sample $Y = (y_1, \ldots, y_n)$ and $T(Y^*)$ denote the M-estimator for the sample $Y^* = Y \cup Z$.

Theorem (Huber, 1983b). Put

$$\sum_{y_i \in Y} \rho(y_i - T(Y)) = -A$$

for some constant $A > 0$. (Note that $n \geq A$.) Then the breakdown point of a redescending M-estimator is given by

$$\epsilon^*(Y, T) = \frac{m^*}{n + m^*}, \qquad (A19)$$

where m^* is an integer satisfying $\lceil A \rceil \leq m^* \leq \lfloor A \rfloor + 1$. ($\lceil A \rceil$ denotes the smallest integer not smaller than A and $\lfloor A \rfloor$ denotes the largest integer not larger than A.) If there is a constant $k < \infty$ such that $\rho(u) \equiv 0$ for $|u| > k$ then $m^* = \lceil A \rceil$.

Remark. Huber conjectures that if $\rho(u)$ is strictly negative for all finite u, then $m^* = \lfloor A \rfloor + 1$.

Proof. First consider the case in which $m < A$, where m is the number of points in Z; then $T(Y^*)$ does not break down. Clearly

$$\sum_{y_i^* \in Y^*} \rho(y_i^* - T(Y)) \leq A \tag{A20}$$

since

$$\sum_{y_i \in Y} \rho(y_i - T(Y)) \leq A$$

by assumption. Note that equality in (A20) obtains iff $\rho(z_i) = 0$ for all $z_i \in Z$. Choose some $\delta > 0$ such that $m + n\delta < A$. Then let k satisfy $\rho(u) \geq -\delta$ for $|u| \geq k$, and let t be any real number such that $|y_i - t| \geq k$ for all $y_i \in Y$. If for arbitrarily large k, t minimizes $\sum \rho(y_i^* - t)$ for the $y_i^* \in Y^*$, then the estimator $T(Y^*)$ will break down. However

$$\sum_{y_i \in Y} \rho(y_i - t) \geq -n\delta$$

by the choice of t. Similarly

$$\sum_{z_i \in Z} \rho(z_i - t) \geq -m\,.$$

Hence,

$$\sum_{y_i^* \in Y^*} \rho(y_i^* - t) \geq -(n\delta + m)\,. \tag{A21}$$

But this implies that

$$\sum_{y_i^* \in Y^*} \rho(y_i^* - T(Y)) < \sum_{y_i^* \in Y^*} \rho(y_i^* - t)\,,$$

that is, the left-hand side of (A20) is strictly less than the left-hand side of (A21). Then $T(Y^*)$ falls within a distance k from any point in Y and so does not break down.

Now consider the case $m > A$. Let $\delta > 0$ satisfy $m - m\delta > A$ and k satisfy $\rho(u) \geq -\delta$ for $|u| \geq k$. Also let all points in $Z = (z_1, \ldots, z_m)$ be equal to $\tilde{z}$, where $\tilde{z}$ is any real number such that $|\tilde{z} - T(Y)| \geq k$, and let t be any real number such that $|\tilde{z} - t| \geq k$. Then clearly

$$\sum_{z_i \in Z} \rho(z_i - t) \geq -m\delta$$

and

$$\sum_{y_i \in Y} \rho(y_i - t) \geq -A$$

Hence,

$$\sum_{y_i^* \in Y^*} \rho((y_i^* - t) \geq -(A + m\delta) \,. \tag{A22}$$

Similarly,

$$\sum_{z_i \in Z} \rho(z_i - \tilde{z}) = -m$$

and

$$\sum_{y_i \in Y} \rho(y_i - \tilde{z}) \leq 0 \,.$$

Hence,

$$\sum_{y_i^* \in Y^*} \rho((y_i^* - \tilde{z}) \leq -m \,. \tag{A23}$$

But this implies that

$$\sum_{y_i^* \in Y^*} \rho(y_i^* - \tilde{z}) < \sum_{y_i^* \in Y^*} \rho(y_i^* - t) \,,$$

that is, the left-hand side of (A23) is strictly less than the left-hand side of (A22). Then $T(Y^*)$ falls within a distance k from $\tilde{z}$. Hence, if $\tilde{z} \to \infty$, breakdown occurs.

Finally, consider the case $m = A$. If $\rho(u) \equiv 0$ for $|u| \geq k$, then (A22) is true if $\delta = 0$ and $\tilde{z}$ is chosen such that $|y - \tilde{z}| \geq k$ for all $y \in Y$. Then

$$\sum_{y_i \in Y} \rho(y_i - T(Y)) = -A$$

and

$$\sum_{z_i \in Z} \rho(z_i - T(Z)) = -A \,,$$

if all points in $Z = (z_1, \ldots, z_m)$ are set equal to $\tilde{z}$. Hence, the M-estimator $T(Y^*)$ has two possible solutions, $T(Y^*) = T(Y)$ and $T(Y^*) = \tilde{z}$. But since the M-estimator T is location equivariant, it cannot choose between the two solutions. Then letting $\tilde{z} \to \infty$ causes one solution to be infinite; hence, T breaks down.

Let m^* be an integer satisfying $\lceil A \rceil \leq m^* \leq \lfloor A \rfloor + 1$. Then the three cases imply that

$$\epsilon^*(Y, T) = \frac{m^*}{n + m^*} \,\blacksquare$$

APPENDIX 4

This appendix sketches a proof of the asymptotic variance of T given in expression (28); for details, see Huber (1981, pp. 38–40). Let F_n be the empirical distribution of a sample drawn from F. Under certain regularity conditions, the influence curve can be written as

$$\mathrm{IC}(y; F, T, S) = \lim_{n\to\infty} \frac{T(F_n) - T(F)}{1/n} \tag{A24}$$

It is often possible to expand T in a Taylor series in terms of the influence curve

$$T(F_n) - T(F) = \int \mathrm{IC}(y; F, T, S)[\mathrm{d}F_n(y) + \mathrm{d}F(y)] + o([\mathrm{IC}(y; F, T, S)]^2) \tag{A25}$$

where $o([\mathrm{IC}(y; F, T, S)]^2)$ refers to second and higher order terms of the influence curve. Recall that for M-estimators, $\mathrm{IC}(y; F, T, S)$ is proportional to ψ. Hence, the integral of $\mathrm{IC}(y; F, T, S)$ over F vanishes

$$\int \mathrm{IC}(y; F, T, S)[\mathrm{d}F_n(y) - \mathrm{d}F(y)] = \int \mathrm{IC}(y; F, T, S)\,\mathrm{d}F_n(y)$$

The integral over the empirical distribution F_n can be reexpressed in a somewhat more familiar form

$$\int \mathrm{IC}(y; F, T, S)\mathrm{d}F_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}(y; F, T, S) \tag{A26}$$

Then substituting (A26) into (A25) and multiplying both sides by $\sqrt{n}$ yields

$$\sqrt{n}\,(T(F_n) - T(F)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathrm{IC}(y; F, T, S) + o([\mathrm{IC}(y; F, T, S)]^2) \tag{A27}$$

Often, the last term of the right-hand side of (A27) can be shown to be negligible with respect to the lower order terms. Then by the central limit theorem, the first term of the right-hand side of (A27) is asymptotically $\mathrm{Gau}(0, A^2(F, T, S))$, where

$$A^2(F, T, S) = \int [\mathrm{IC}(y; F, T, S)]^2 \mathrm{d}F(y) \tag{A28}$$

APPENDIX 5

This appendix compares OLS estimates, bounded influence regression estimates, and Bell regression M-estimates for age-adjusted mortality data in 60 Standard Metropolitan

Statistical Areas. These data have been previously analyzed by Henderson and Velleman (1981) and Krasker and Welsch (1982).

Table A1 presents estimates for the three estimators. The specification follows that used by Krasker and Welsch (1982) and regresses age-adjusted mortality for the 60 metropolitan areas on the percentage nonwhite population, average years of education, population (in thousands) per square mile, precipitation, and the logarithm of sulfur dioxide potential (a pollution indicator). The OLS estimates presented in Table A1 differ slightly from those given by Krasker and Welsch and are due to different treatment of rounding for the logarithm of sulfur dioxide potential.

Krasker and Welsch note that the most noticeable differences between the OLS and BIF estimates occur for the effects of the percentage nonwhite population, population density, and precipitation. Note however that the differences in coefficients are within 1.5 standard errors of one another for these parameters. The BIF and Bell estimates are quite close in magnitude; the most noticeable differences are for the effects of population density and the logarithm of sulfur dioxide potential. Qualitative differences between the three estimators are slight.

NOTES

1. Rousseeuw (1982) describes a highly robust least median of squares estimator that could also be used as a starting estimate.

2. I examined a number of standard diagnostic measures including the diagonals of the "hat" matrix, studentized residuals, Cook's D, DFBETAS and DFFITS (see Belsley, Kuh, and Welsch, 1980, and Cook and Weisberg, 1982b for details and definitions). At various cutoff levels suggested by Belsley, Kuh, and Welsch (1980, p. 28), Huber (1981, p. 162), and Cook and Weisberg (1982b, pp. 25-28, 118, 151–156), none of the procedures identified *any* of the observations in the outlying of cluster points as potential leverage points. Similarly, inspecting the 10 largest scores for each measure identified at most four (diagonals of the hat matrix) and more usually none (the remaining measures) of the 20 observations in the outlying cluster. In contrast, the 20 largest residuals obtained from the highly robust M-estimator belong to the 20 observations in the outlying cluster; the smallest such residual is −24.6, while the largest residual for the remaining 30 observations is 0.486.

3. Huber (1977, 1981) and Hoaglin, Mosteller, and Tukey (1983) present thorough overviews of robust statistics. Goodall (1983) and Iglewicz (1983) give excellent introductions to the theory of M-estimation and robust tests, respectively. Early seminal works include papers by Tukey (1960), Huber (1964), and Hampel (1971, 1974a). Andrews et al. (1972) report results from the 1972 Princeton Robustness Study, an early Monte Carlo study of the small sample properties of over 60 estimators. Mosteller and Tukey (1977) provide a nontechnical introduction to many concepts and issues. See also the review articles of Huber (1972), Hampel (1973, 1975), Bickel (1976), and Hogg (1979). For interesting historical details, see Stigler (1973).

4. Distributions with heavier tails than the Gaussian distribution appear to be common in practice; see, for example, Student (1927), Daniel and Wood (1980, Ch. 5), Hampel (1973), Mosteller and Tukey (1977, Ch. 1), Agee and Turner (1979), Hogg (1979), Kleiner, Martin, and Thomson (1979), Huber (1981, p. 91), and Rocke, Downs, and Rocke (1982). For historical examples drawn from astronomical observations, see works cited in Stigler (1973). It should be noted that classical estimators often have good properties for distributions with slightly lighter tails than a Gaussian distribution. Such distributions typically

arise from somewhat artificial situations, for example, distributions of standardized test scores (Hogg, 1974).

5. Section 5 describes estimators of spread that are more robust than either s_n or d_n.

6. Some rather complex adaptive nonparametric estimators can achieve full efficiency. These estimators, proposed by Stein (1956), assume a symmetric underlying distributions only and use the sample distribution to approximate aspects of the population distribution. See Bickel (1982) for a recent overview and Donoho and Huber (1982) for some breakdown calculations. Hogg (1974) reviews some simpler adaptive estimators.

7. A number of other estimation procedures are commonly used, particularly one-step solutions to (14) (see Bickel, 1975) and simultaneous estimation of location and scale (see Huber, 1977, 1981). Estimation is discussed in greater detail in Section 5.

8. The definition of $\text{sign}(0) = 1$ is somewhat arbitrary since the derivative of $|u|$ does not exist at $u = 0$. For consistency, I define the sample median to be the larger of the two middle observations when the sample size n is even.

9. More formally, the delta function $\delta(u)$ may be defined as the limit of any sequence of functions $\{\delta_k(u)\}$ such that $\int_{-\infty}^{\infty} g(u)\, \delta_k(u)\, du = g(0)$ as $k \to \infty$, where $g(u)$ is any bounded, integrable, and continuous function.

10. Table 3 presents the customary definition of the Huber M-estimator, which is given in terms of the tuning constant c and not the tuning constant k in the denominator of $u \equiv (y - \theta)/kS_n$. The Hampel M-estimator is similarly defined but has three tuning constants a, b, and c. Hence $k \equiv 1$ for these two estimators.

11. For example, Freedman and Diaconis construct certain symmetric but multimodal densities for which the bisquare M-estimator, defined by minimizing $\sum \rho(u)$, is inconsistent for $k < 5.4$; they note, however, that a consistent estimator is obtained for these densities if $k \geq 5.4$ or the M-estimator is defined by the solution of (14) or (15) closest to the median. A practical implication of these results is that k for redescending M-estimators should not be chosen too small or, equivalently, ψ should not redescend too quickly. Note that small values of k are typically avoided in order not to excessively degrade efficiency for the Gaussian distribution.

12. Replacing the continuous distribution F by the finite sample distribution F_{n-1} and ϵ by $1/n$ in expressions (23) and (24) yields the sensitivity curve (Goodall, 1983). The

jackknife and bootstrap (see, for example, Efron, 1982) are closely related to the influence curve.

13. The definition of resistance presented here accords extremely well with a more technical concept called *qualitative robustness*, where qualitative robustness is defined either by continuity in the weak-star topology or the equicontinuity of sequences of estimators. For discussion and details, see Hampel (1971) and Huber (1981).

14. The small sample variances for two bisquare estimators with tuning constants 6.0 and 8.8 are taken from Goodall (1983). Small sample variances for the bisquare estimator with tuning constant 6.4 and the Bell estimator are taken from Bell (1980). Remaining small sample results are taken from results of the Princeton Robustness Study reported in Andrews et al. (1972). The results reported in Goodall are a continuation of the Princeton Robustness study and were obtained using the same Monte Carlo data of 640 to 1000 samples of size 20; hence, results should be comparable for these two studies. Bell uses substantially larger numbers of samples in his Monte Carlo simulations (10,000, 20,000, and 100,000 samples of size 20 for the Gaussian, 1WG, and slash distributions, respectively) to obtain greater accuracy. Hence, some care is needed in comparing the results reported in Table 4.

15. Finite sample variances for the subsample mean and slash Pitman estimators are taken from values reported in Goodall (1983).

16. The differences between the small sample and asymptotic efficiencies for some estimators are surprising. One possibility is the relatively small Monte Carlo samples used in some of the studies. This may account for discrepancies between the small sample and asymptotic performance for the slash distribution for some estimators.

17. I would like to thank Bradley Efron for suggesting the use of Johnson's t-statistic.

18. A number of other possible approaches have been suggested in the literature. Boos (1980) constructs approximate confidence intervals; however, this approach is limited to non-redescending ψ. Another approach involves estimating the appropriate degrees of freedom using approximations derived from asymptotic theory. Shoemaker and Hettsmansperger (1982) derive one such expression and show that A_n^2 has a distribution close to a chi-square distribution using the estimated degrees of freedom; however they do not examine the performance of the statistic in (32) using the estimated degrees of freedom. Perhaps most promising are sample reuse methods such as the bootstrap (see Efron, 1982). How-

ever, such methods are unusually computer intensive, which may pose practical difficulties. Rocke and Downs (1981) report some preliminary findings on the performance of A_n^2 and of jackknife and bootstrap estimators for the variances of several distributions. Iglewicz (1983) examines both A_n^2 and simpler estimators of spread; he also provides a broad overview of the construction of robust tests and confidence intervals.

19. Some preliminary evidence (Rocke and Downs, 1981) indicates that the simple tests based on the t-like statistic in (32) perform well for symmetric distributions but less well for asymmetric distributions. Johnson (1978) and Efron (1981) compare the performance of the classical Student t, t^*, and a bootstrap t for the mean under sampling from a highly skewed asymmetric distribution in Monte Carlo simulations. Johnson's simulations indicate that t^* performed far better than the classical student t while Efron's simulations indicate a close agreement between the performance of t^* and the theoretically superior bootstrap t, which requires far greater computational resources. These results provide heuristic support for the use of t^*. Note, in particular, that small samples drawn from heavy tailed symmetric distributions may have large random asymmetries.

20. An exception is the performance of the Bell M-estimator for the slash distribution, where the test with $n-1$ df performs better than the test with $0.7(n-1)$ df . (The bisquare estimator shows a similar but less marked trend.) One explanation is that since both the bisquare and Bell estimators are defined by redescending ψ functions, they have a somewhat slower rate of convergence to their asymptotic performance for light tailed distributions relative to their rate of convergence for heavier tailed distributions. Hence, the $0.7(n-1)$ df , needed for the Gaussian and contaminated Gaussian distributions, overcompensates somewhat for the slash distribution.

21. For additional discussion of robust estimation of linear models, see Andrews (1974), Huber (1973), Hill and Holland (1977), Hogg (1979), Carroll (1980), and Huber (1981).

22. Krasker and Welsch (1982) propose an alternative to the M-estimators of regression examined in this section that is intended to deal specifically the problems caused by leverage points. The "bounded influence" regression estimator is chosen to minimize the asymptotic variance for Gaussian error distributions subject to a bound on the asymptotic gross error sensitivity and incorporates some standard diagnostic measures into its weighting function. Although not designed specifically to deal with leverage, redescending M-estimators

of regression nevertheless have excellent breakdown (typically $\epsilon^* > .40$) against severe contamination, including contamination by leverage points (Huber, 1983b; Donoho, 1984). In contrast, the Krasker–Welsch estimator has poorer breakdown ($\epsilon^* = 1/(p+1)$, where p is the rank of the matrix $\mathbf{X}$; see Maronna, Yohai, and Bustos, 1979). Huber (1983a, with discussion) discusses the Krasker–Welsch proposal in detail. Appendix 5 compares OLS estimates, bounded influence regression estimates, and Bell regression M-estimates for data analyzed by Krasker and Welsch (1982).

REFERENCES

AGEE, W. S., AND TURNER, R. H.

1979 "Application of robust regression to trajectory data reduction." In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics.* New York: Academic Press.

ANDREWS, D. F.

1974 "A robust method for linear regression." *Technometrics* **16**(4): 523–531.

1979 "The robustness of residual displays." In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics.* New York: Academic Press.

ANDREWS, D. F., AND PREGIBON, D.

1978 "Finding the outliers that matter." *Journal of the Royal Statistical Society* **B40**(1): 85–93.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., AND TUKEY, J. W.

1972 *Robust Estimates of Location: Survey and Advances.* Princeton: Princeton University Press.

ANSCOMBE, F. J.

1960 "Rejection of outliers." *Technometrics* **2**(2): 123–147.

ATKINSON, A. C.

1982 "Regression diagnostics, transformations and constructed variables." *Journal of the Royal Statistical Society* **B44**(1): 1–36.

BEATON, A. E., AND TUKEY, J. W.

1974 "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data." *Technometrics* **16**(2): 147–185.

BELL, R. M.

1980 "An adaptive choice of the scale parameter for M-estimators." Technical Report No. 3., Department of Statistics, Stanford University.

BELSLEY, D. A., KUH, E., AND WELSCH, R. E.

1980 *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

BENJAMINI, Y.

1983 "Is the t test really conservative when the parent distribution is long-tailed?" *Journal of the American Statistical Association* **78**(383): 645–654.

BICKEL, P. J.

1975 "One-step Huber estimates in the linear model." *Journal of the American Statistical Society* **70**(350): 428–434.

1976 "Another look at robustness: a review of reviews and some new developments." *Scandinavian Journal of Statistics* **3**(4): 145–168.

1982 "On adaptive estimation." *Annals of Statistics* **10**(3): 647–671.

BOOS, D. D.

1980 "A new method for constructing approximate confidence intervals from M estimates." *Journal of the American Statistical Association* **75**(369): 142–145.

BORUS, M. E., AND SANTOS, R.

1983 "The youth population." In M. E. Borus (ed.), *Tommorrow's Workers.* Lexington, MA: Lexington Books.

BYRD, R. H., AND PYNE, D. A.

1979 "Some results on the convergence of the iteratively reweighted least squares algorithm for robust regression." *Proceedings of the American Statistical Association, Statistical Computing Section.*

CARROLL, R. J.

1980 "Robust methods for factorial experiments with outliers." *Applied Statistics* **29**(3): 246–251.

COLEMAN, D., HOLLAND, P., KADEN, N., KLEMA, V., AND PETERS, S. C.

1980 "A system of subroutines for iteratively reweighted least squares computations." *ACM Transactions of Mathematical Software* **6**(3): 327–336.

COLLINS, J. R.

1976 "Robust estimation of a location parameter in the presence of asymmetry." *Annals of Statistics* **4**(1): 68–85.

COOK, R. D.

1977 "Detection of influential observations in linear regression." *Technometrics* **19**(1): 15–18.

COOK, R. D., AND WEISBERG, S.

1982a "Criticism and influence in linear regression." In S. Leinhardt (ed.), *Sociological Methodology 1982.* San Francisco: Jossey-Bass.

1982b *Residuals and Influence in Regression.* New York: Chapman and Hall.

DANIEL, C., AND WOOD, F. S.

1980 *Fitting Equations to Data.* 2nd ed. New York: Wiley.

DONOHO, D. L.

1982 "Breakdown properties of multivariate location estimators." Department of Statistics, Harvard University.

1984 Personal communication.

DONOHO, D. L., AND HUBER, P. J.

1983 "The notion of breakdown point." In P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr. (eds.), *A Festschrift for Erich L. Lehmann.* Belmont, CA: Wadsworth.

EFRON, B.

1981 "Nonparametric standard errors and confidence intervals." *Canadian Journal of Statistics* **9**(2): 139–172.

1982 *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS Regional Conference Series in Applied Mathematics 38. Philadelphia: Society for Industrial and Applied Mathematics.

FISHER, R. A.

1920 "A mathematical examination of the methods of determining the accuracy of an observation by the mean error and the mean square error." *Monthly Notices of the Royal Astronomical Society* **80**: 758–770.

FREEDMAN, D. A., AND DIACONIS, P.

1982 "On inconsistent M-estimators." *Annals of Statistics* **10**(2): 454–461.

FRIEDMAN, J. H., AND STUETZLE, W.

1982 "Projection pursuit methods for data analysis." In R. L. Launer and A. F. Siegel (eds.), *Modern Data Analysis.* New York: Academic Press.

GASKO, M., AND DONOHO, D. L.

1982 "Influential observation in data analysis." *Proceedings of the American Statistical Association, Business and Economic Statistics Section.*

GOODALL, C.

1983 "M-estimators of location: an outline of the theory." In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

GROSS, A. M.

1976 "Confidence interval robustness with long-tailed symmetric distributions." *Journal of the American Statistical Association* **71**(354): 409–416.

HAMPEL, F. R.

1971 "A general qualitative definition of robustness." *Annals of Mathematical Statistics* **42**(6): 1887–1896.

1973 "Robust estimation: a condensed partial survey." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **27**: 87–104.

1974a "The influence curve and its role in robust estimation." *Journal of the American Statistical Association* **69**(346): 383–393.

1974b "Rejection rules and robust estimates of location: an analysis of some monte carlo results." *Proceedings of the European Meeting of Statisticians and 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes.* Prague, Czechoslovakia, 1974.

1975 "Beyond location parameters: robust concepts and methods." *Proceedings of the Prague Symposium on Asymptotic Statistics.* Prague, Czechoslovakia, 1975.

HENDERSON, H. V., AND VELLEMAN, P. F.

1981 "Building multiple regression models interactively." *Biometrics* **37**(2): 391–411.

HILL, R. W., AND HOLLAND, P. W.

1977 "Two robust alternatives to least-square regression." *Journal of the American Statistical Association* **72**(360): 828–833.

HOAGLIN, D. C., AND WELSCH, R. E.

1978 "The hat matrix in regression and ANOVA." *American Statistician* **32**(1): 17–22 and Corrigenda **32**(4): 146.

HOAGLIN, D. C., MOSTELLER, F., AND TUKEY, J. W.

1983 *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

HOGG, R. V.

1974 "Adaptive robust procedures." *Journal of the American Statistical Association* **69**(348): 909–927.

1979 "Statisticalrobustness: one view of its use in applications today." *American Statistician* **33**(3): 108–115.

HOLLAND, P. W., AND WELSCH, R. E.

1977 "Robust regression using iteratively reweighted least-squares." *Communications in Statistics* **A6**(9): 813–827.

HUBER, P. J.

1964 "Robust estimation of a location parameter." *Annals of Mathematical Statistics* **35**(1): 73–101.

1972 "Robust statistics: a review." *Annals of Mathematical Statistics* **43**(4): 1041–1067.

1973 "Robust regression: asymptotics, conjectures and monte carlo." *Annals of Statistics* **1**(5): 799–821.

1977 *Robust Statistical Procedures.* CBMS Regional Conference Series in Applied Mathematics 27. Philadelphia: Society for Industrial and Applied Mathematics.

1981 *Robust Statistics.* New York: Wiley.

1983a "Minimax aspects of bounded-influence regression." *Journal of the American Statistical Association* **78**(381): 66–80.

1983b "Finite sample breakdown of M- and P-estimators." Unpublished manuscript, Department of Statistics, Harvard University.

IGLEWICZ, B.

1983 "Robust scale estimators and confidence intervals for location." In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

JOHNS, M. V.

1979 "Robust Pitman-like estimators." In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics.* New York: Academic Press.

JOHNSON, N. J.

1978 "Modified *t* tests and confidence intervals for asymmetric populations." *Journal of the American Statistical Association* **73**(363): 536–544.

KAFADAR, K.

1982 "A biweight approach to the one-sample problem." *Journal of the American Statistical Association* **77**(378): 416–424.

KLEINER, B., MARTIN, R. D., AND THOMSON, D. J.

1979 "Robust estimation of power spectra." *Journal of the Royal Statistical Society* **B41**(3): 313–351.

KRASKER, W. S., AND WELSCH, R. E.

1982 "Efficient bounded-influence regression estimation." *Journal of the American Statistical Association* **77**(379): 595–604.

LEINHARDT, S., AND WASSERMAN, S. S.

1978 "Exploratory data analysis: an introduction to selected methods." *Sociological Methodology 1979.* San Francisco: Jossey-Bass.

MARONNA, R. A., YOHAI, V. T., AND BUSTOS, O. H.

1979 "Bias- and Efficiency-Robustness of General M-estimators for Regression with Random Carriers." In T. A. Gasser and M. Rosenblatt (eds.), *Smoothing Techniques for Curve Estimation.* New York: Springer Verlag.

MARTINEZ, J., AND IGLEWICZ, B.

1981 "A simple procedure for finding t-type robust confidence intervals." *Proceedings of the American Statistical Association, Statistical Computing Section.*

MOSTELLER, F., AND TUKEY, J. W.

1977 *Data Analysis and Regression.* Reading, MA: Addison-Wesley.

NORTON, R. D.

1979 *City Life-Cycles and American Urban Policy.* New York: Academic Press.

RELLES, D. A., AND ROGERS, W. H.

1977 "Statisticians are fairly robust estimators of location." *Journal of the American Statistical Association* **72**(357): 107–111.

ROCKE, D. M., AND DOWNS, G. W.

1981 "Estimating the variances of robust estimators of location: influence curve, jackknife and bootstrap." *Communications in Statistics* **B10**(3): 221–248.

ROCKE, D. M., DOWNS, G. W., AND ROCKE, A. J.

1982 "Are robust estimators really necessary?" *Technometrics* **24**(2): 95–101.

ROGERS, W. H., AND TUKEY, J. W.

1972 "Understanding some long-tailed symmetrical distributions." *Statistica Neerlandica* **26**(3): 211–226.

ROUSSEEUW, P. J.

1982 "Least median of squares regression." Unpublished manuscript, Centrum voor Statistieken Operationeel Onderzoek, Vrije Universiteit, Brussels, Belgium.

SHOEMAKER, L. H., AND HETTSMANSPERGER, T. P.

1982 "Robust estimates and tests for the one- and two-sample scale models." *Biometrika* **69**(1): 47–53.

SHORACK, G. R.

1976 "Robust studentization of location estimates." *Statistica Neerlandica* **30**(3): 119–141.

SIEGEL, A. F.

1982 "Robust regression using repeated medians." *Biometrika* **69**(1): 242–244.

STEIN, C.

1956 "Efficient nonparametric testing and estimation." In J. Neyman (ed.), *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1.* Berkeley: University of California Press.

STIGLER, S. M.

1973 "Simon Newcomb, Percy Daniell and the history of robust estimation 1885–1920." *Journal of the American Statistical Association* **68**(344): 872–879.

STOTO, M. A., AND EMERSON, J. D.

1983 "Power transformations for data analysis." In S. Leinhardt (ed.), *Sociological Methodology 1983–1984.* San Francisco: Jossey-Bass.

STUDENT.

1927 "Errors of routine analysis." *Biometrika* **19**: 151-164.

TUKEY, J. W.

1960 "A survey of sampling from contaminated distributions." In I. Olkin et al. (eds.), *Contributions to Probability and Statistics.* Stanford, CA: Stanford University Press.

1979a "Robust techniques for the user." In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics.* New York: Academic Press.

1979b "Study of robustness by simulation: particularly improvement by adjustment and combination." In R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics.* New York: Academic Press.

Figure Captions

Figure 1: Simple regression for hypothetical data contaminated by a cluster of leverage points. Solid line: OLS estimate; dotted line: M-estimate started from the OLS estimate; dashed line: M-estimate started from the repeated median estimate.

Figure 2: Normalized densities $f(y)$ for the Gaussian, logistic, double exponential, and Cauchy distributions defined in Table 2. The location parameter $\theta = 0$. Scale is assumed known and is chosen such that $f(0) = 1/\sqrt{2\pi}$. To better compare the extreme tail behavior of the densities, the right-hand-side illustrates $f(y)$ for $3 \leq y \leq 12$; $f(y)$ is multiplied by a factor of 10 in this range. Solid line: Gaussian distribution; dot-dashed line: logistic distribution; dashed line: double exponential distribution; dotted line: Cauchy distribution.

Figure 3: $\psi(u)$ corresponding to the ML-estimators for the Gaussian, logistic, double exponential, and Cauchy densities in Figure 2. Solid line: Gaussian distribution; dot-dashed line: logistic distribution; dashed line: double exponential distribution; dotted line: Cauchy distribution.

Figure 4: ρ, ψ, and ϕ for the mean and three hypothetical points, $\hat{u}_1 = -3$, $\hat{u}_2 = 1$, and $\hat{u}_3 = 2$, corresponding to the deviations $u_i = y_i - \theta$ of three observations y_1, y_2, and y_3 obtained by setting θ equal to the sample mean.

Figure 5: $\rho(u)$ for the median, Huber, Hampel, sine, bisquare, and Bell estimators.

Figure 6: $\psi(u)$ for the median, Huber, Hampel, sine, bisquare, and Bell estimators.

Figure 7: $\phi(u)$ for the median, Huber, Hampel, sine, bisquare, and Bell estimators.

Figure 8: Scatterplot of the percentage growth of employment in the manufacturing sector and the percentage growth of employment in the service sector, 30 largest standard metropolitan statistical areas. Solid line: OLS estimate; dashed line: Bell/RM estimate.

Figure 9: Gaussian probability plot of the expected Gaussian quantiles versus residuals obtained from the OLS and Bell/RM estimators, full sample ($n = 30$). Solid line: OLS residuals; dashed line: Bell/RM residuals.

Figure 10: Stem and leaf display for untransformed hourly wages in 1979 dollars for 19 year old white males employed and out of school, full sample ($n = 120$). The outlying observation has wages of \$404/hour.

Figure 11: Partial regression plot for collective bargaining in the workplace, restricted sample ($n = 119$). The residuals $e(x_{\text{colbar}\cdot\text{rest}})$ and $e(y_{\cdot\text{rest}})$ are plotted along the x- and y-axes, respectively.

Figure 12: Partial regression plot for urban residence, restricted sample ($n = 119$). The residuals $e(x_{\text{urban}\cdot\text{rest}})$ and $e(y_{\cdot\text{rest}})$ are plotted along the x- and y-axes, respectively.

Table 1: Asymptotic relative efficiency of the mean absolute deviation d_n to the mean square deviation s_n for an ϵ-contaminated Gaussian distribution at scale 3.

ϵ	ARE(ϵ)
0	0.876
.001	0.948
.002	1.016
.005	1.198
.01	1.439
.02	1.751
.05	2.035
.10	1.903
.2	1.510
.5	1.017
1	0.876

Table 2: The densities $f(y;\theta)$ for the Gaussian, logistic, double exponential, and Cauchy distributions and the functions ρ and ψ for the corresponding ML-estimators.

Distribution	$f(y;\theta)$	$\rho(u)$	$\psi(u)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp[-(y-\theta)^2/2]$	$\frac{1}{2}\log 2\pi + \frac{1}{2}u^2$	u
logistic	$\frac{1}{\sqrt{2\pi}}\cosh\left[\sqrt{\frac{2}{\pi}}\,(y-\theta)\right]$	$\frac{1}{2}\log 2\pi + 2\log\cosh\left(\sqrt{\frac{2}{\pi}}\,u\right)$	$2\sqrt{\frac{2}{\pi}}\,\tanh\left(\sqrt{\frac{2}{\pi}}\,u\right)$
double exponential	$\frac{1}{\sqrt{2\pi}}\exp\left(\sqrt{\frac{2}{\pi}}\,\lvert y-\theta\rvert\right)$	$\frac{1}{2}\log 2\pi + \sqrt{\frac{2}{\pi}}\,u$	$\sqrt{\frac{2}{\pi}}\,\mathrm{sign}(u)$
Cauchy	$\frac{1}{\sqrt{2\pi}}\left[1+\frac{\pi}{2}(y-\theta)^2\right]^{-1}$	$\frac{1}{2}\log\frac{2}{\pi} + \log\left(1+\frac{\pi}{2}u^2\right)$	$\frac{2\pi u}{2+\pi u^2}$

Table 3: Some common M-estimators of location.[a]

Estimator	$\rho(u)$	$\psi(u)$	$\phi(u)$	Range of u
mean	$\frac{1}{2}u^2$	u	1	$\|u\| < \infty$
median	$\|u\|$	$\text{sign}(u)$	$\delta(u)$	$\|u\| < \infty$
Huber,	$\frac{1}{2}u^2$	u	1	$\|u\| \leq c$
$c > 0$,				
$k \equiv 1$	$c\,\|u\| - \frac{1}{2}c^2$	$c\,\text{sign}(u)$	0	$\|u\| > c$
Hampel,	$\frac{1}{2}u^2$	u	1	$\|u\| \leq a$
$0 < a \leq b \leq c$,				
$k \equiv 1$	$a\,\|u\| - \frac{1}{2}a^2$	$a\,\text{sign}(u)$	0	$a < \|u\| \leq b$
	$ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}\left[1 - \left(\frac{c-\|u\|}{c-b}\right)^2\right]$	$a\frac{c-\|u\|}{c-b}\,\text{sign}(u)$	$-\frac{a}{c-b}\,\text{sign}(u)$	$b < \|u\| \leq c$
	$ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}$	0	0	$\|u\| > c$
Andrews' sine,	$\frac{1-\cos(\pi u)}{\pi^2}$	$\frac{1}{\pi}\sin(\pi u)$	$\cos(\pi u)$	$\|u\| \leq 1$
$k > 0$				
	$2/\pi^2$	0	0	$\|u\| > 1$
Tukey's	$\frac{1}{6}[1-(1-u^2)^3]$	$u(1-u^2)^2$	$(1-u^2)(1-5u^2)$	$\|u\| \leq 1$
bisquare,				
$k > 0$	$1/6$	0	0	$\|u\| > 1$
Bell,	$\frac{5}{4}\left[1-\left(1+\frac{u^2}{5}\right)^{-2}\right]$	$u\left(1+\frac{u^2}{5}\right)^{-3}$	$(1-u^2)\left(1+\frac{u^2}{5}\right)^{-4}$	$\|u\| < \infty$
$k > 0$				

[a] Tuning constants are denoted by a, b, c, and k. The function $\rho(u)$ is defined up to arbitrary additive and multiplicative constants. Following Holland and Welsch (1977) and Goodall (1983), constants are chosen such that $\rho(0) = 0$ and $\phi(0) = 1$.

Table 4: Finite sample ($n = 20$) relative efficiencies for the M-estimators in Table 3.[a]

Estimator	Tuning	Scale	Gaussian	1WG	Slash
	constant	parameter			
mean			100.0%	16.2%	0.0%
			(1.000)	(6.485)	(12951.48)
median			66.8%	67.7%	84.1%
			(1.498)	(1.555)	(6.60)
Huber	1.5	Normed	95.2%	86.2%	63.5%
		MAD	(1.050)	(1.222)	(8.75)
	2.0		98.1%	82.7%	52.8%
			(1.019)	(1.273)	(10.52)
Hampel	1.2, 3.5, 8.0	MAD	83.0%	86.0%	89.1%
			(1.205)	(1.225)	(6.23)
	1.7, 3.4, 8.5		88.5%	90.3%	82.1%
			(1.130)	(1.166)	(6.76)
	2.5, 4.5, 9.5		95.6%	93.4%	69.3%
			(1.046)	(1.127)	(8.01)
sine	2.1π	MAD	93.5%	93.1%	71.7%
			(1.070)	(1.131)	(7.74)
bisquare	6.0	MAD	86.4%	89.4%	81.8%
			(1.158)	(1.177)	(6.79)
	6.4		89.0%	88.9%	86.9%
			(1.123)	(1.184)	(6.39)
	8.8		96.1%	93.6%	68.4%
			(1.041)	(1.125)	(8.12)
Bell	1/0.35	MAD	90.6%	89.5%	88.4%
			(1.103)	(1.177)	(6.28)
Reference			100.0%	100.0%	100.0%
			(1.000)	(1.052)	(5.552)

[a] Variances (times n) are reported in parentheses. Reference estimators are the Pitman estimators for the Gaussian and slash distributions and the subsample mean for the 1WG.

Table 5: Asymptotic relative efficiencies for the M-estimators in Table 3.[a]

Estimator	Tuning	Scale	Gaussian	5%C10	Slash
	constant	parameter			
mean			100.0%	17.5%	0.0%
			(1.000)	(5.950)	(∞)
median			63.7%	60.3%	77.1%
			(1.571)	(1.722)	(6.283)
Huber	1.5	Normed	96.4%	84.6%	65.7%
		MAD	(1.037)	(1.228)	(7.379)
	2.0		99.0%	81.6%	54.3%
			(1.010)	(1.273)	(8.925)
Hampel	1.2, 3.5, 8.0	MAD	85.8%	84.1%	95.9%
			(1.166)	(1.235)	(5.053)
	1.7, 3.4, 8.5		91.6%	89.3%	89.5%
			(1.092)	(1.164)	(5.413)
	2.5, 4.5, 9.5		97.5%	92.7%	75.3%
			(1.025)	(1.121)	(6.441)
sine	2.1π	MAD	96.0%	92.6%	82.5%
			(1.042)	(1.122)	(5.879)
bisquare	6.0	MAD	91.4%	89.8%	89.6%
			(1.094)	(1.156)	(5.412)
	6.4		93.2%	91.1%	87.5%
			(1.073)	(1.141)	(5.538)
	8.8		98.0%	92.8%	75.2%
			(1.020)	(1.120)	(6.443)
Bell	1/0.35	MAD	93.2%	90.3%	88.5%
			(1.073)	(1.151)	(5.475)
Reference			100.0%	100.0%	100.0%
			(1.000)	(1.039)	(4.847)

[a] Asymptotic variances are reported in parentheses. Reference estimators are the ML-estimators.

Table 6: Four estimators with high finite sample and asymptotic triefficiencies.

Finite sample ($n = 20$) results			Asymptotic results		
Estimator	Tuning constant	Triefficiency	Estimator	Tuning constant	Triefficiency
Bell	1/0.35	88.4%	bisquare	6.0	89.6%
bisquare	6.4	86.9%	Hampel	1.7, 3.4, 8.5	89.3%
Hampel	1.2, 3.5, 8.0	83.0%	Bell	1/0.35	88.5%
Hampel	1.7, 3.4, 8.5	82.1%	bisquare	6.4	87.5%

Table 7: Gross error sensitivity for the M-estimators in Table 3.

Estimator	Tuning constant	Scale parameter	Gaussian	5%C10	Slash
mean			∞	∞	∞
median			1.253	1.312	2.507
Huber	1.5	Normed MAD	1.731	1.909	6.448
	2.0		2.095	2.307	7.153
Hampel	1.2, 3.5, 8.0	MAD	1.403	2.207	3.254
	1.7, 3.4, 8.5		1.547	2.456	3.843
	2.5, 4.5, 9.5		1.859	3.002	4.893
sine	2.1π	MAD	1.817	1.967	4.709
bisquare	6.0	MAD	1.680	1.796	4.728
	6.4		1.713	1.840	4.871
	8.8		2.018	2.200	5.827
Bell	1/0.35	MAD	1.642	1.763	4.065

Table 8: Observed level of α based on 5,000 samples of size $n = 20$ drawn from a Gaussian distribution by nominal level α, estimator, and test (t and t^*).[a]

Location	Test	Degrees of	Nominal level α			
estimator	statistic	freedom[b]	.15	.10	.05	.01
Mean	t	19	.158	.105	.0520	.0102
	t^*	19	.157	.104	.0530	.0098
bisquare,	t	19	.163	.109	.0598	.0156
$k = 6.4$	t^*	19	.161	.109	.0584	.0144
	t	13	.155	.101	.0530	.0114
	t^*	13	.153	.100	.0530	.0118
Bell,	t	19	.161	.107	.0596	.0142
$k = 1/0.35$	t^*	19	.158	.106	.0576	.0128
	t	13	.153	.100	.0510	.0104
	t^*	13	.149	.099	.0504	.0100

[a] The test statistics t and t^* are defined in expressions (32) and (34), respectively.

[b] See text.

Table 9: Observed level of α based on 5,000 samples of size $n = 20$ drawn from a 5% Contaminated Gaussian distribution by nominal level α, estimator, and test (t and t^*).[a]

Location	Test	Degrees of	Nominal level α			
estimator	statistic	freedom[b]	.15	.10	.05	.01
Mean	t	19	.131	.074	.0304	.0042
	t^*	19	.230	.148	.0712	.0082
bisquare,	t	19	.160	.113	.0614	.0144
$k = 6.4$	t^*	19	.158	.113	.0592	.0136
	t	13	.152	.105	.0530	.0090
	t^*	13	.152	.104	.0526	.0098
Bell,	t	19	.157	.108	.0594	.0128
$k = 1/0.35$	t^*	19	.155	.108	.0572	.0130
	t	13	.150	.102	.0506	.0090
	t^*	13	.149	.102	.0498	.0086

[a] See notes to Table 8.

Table 10: Observed level of α based on 20,000 samples of size $n = 20$ drawn from a Slash distribution by nominal level α, estimator, and test (t and t^*).[a]

Location	Test	Degrees of	Nominal level α			
estimator	statistic	freedom[b]	.15	.10	.05	.01
Mean	t	19	.113	.060	.0207	.0015
	t^*	19	.270	.172	.0771	.0127
bisquare,	t	19	.155	.105	.0524	.0102
$k = 6.4$	t^*	19	.161	.111	.0583	.0117
	t	13	.148	.096	.0467	.0071
	t^*	13	.154	.104	.0514	.0090
Bell,	t	19	.150	.101	.0492	.0092
$k = 1/0.35$	t^*	19	.154	.105	.0535	.0096
	t	13	.142	.094	.0438	.0060
	t^*	13	.147	.097	.0467	.0071

[a] See notes to Table 8.

Table 11: OLS and Bell estimates for regression of the percentage growth in employment in the manufacturing sector on the percentage growth in employment in the service sector, 30 largest standard metropolitan statistical areas.

	Full Sample ($n = 30$)			Restricted Sample ($n = 29$)	
	OLS	Bell/RM[a]	Bell/OLS[b]	OLS	Bell/RM[a]
Growth in manufacturing	0.44[c]	0.66	0.43	0.74	0.66
	(0.04)[d]	(0.02)	(0.02)	(0.06)	(0.04)
	12.42[e]	16.67	14.84	12.17	12.02
Intercept	61.03	47.86	56.95	45.85	47.89
	(6.49)	(3.27)	(4.47)	(5.40)	(3.74)
	9.40	11.44	11.44	8.49	10.39

[a] Bell estimate started from the robust repeated median estimate.

[b] Bell estimate started from the nonrobust OLS estimate.

[c] Parameter estimates.

[d] Standard errors in parentheses.

[e] t-values (Student t on $n-1$ df and t^* on the $0.7(n-1)$ df for the OLS and Bell estimators, respectively).

Table 12: Seven largest residuals for the OLS and Bell/RM estimates, full (n = 30) sample.[a]

	1	2	3	4	5	6	7
OLS	San Diego 93.37	Houston 72.95	Atlanta 56.82	Phoenix −49.02	New York -44.03	Kansas City −36.66	Pittsburgh −36.26
Bell	Phoenix −221.93 (0.0004)	San Diego 61.47 (0.1335)	Houston 55.82 (0.1748)	Atlanta 45.87 (0.2777)	Dallas −39.93 (0.3617)	Kansas City −34.38 (0.4560)	New York −28.95 (0.5615)

[a] Weights for the Bell/RM estimator are reported in parentheses. Small weights correspond to large residuals.

Table 13: OLS and Bell estimates for determinants of log transformed wages for 19 year old white males employed and out of school.

	Full Sample ($n = 120$)		Restricted Sample ($n = 119$)	
	OLS	Bell	OLS	Bell
Ability score	0.050[a]	0.335**	0.358**	0.334**
	(0.262)[b]	(0.146)	(0.155)	(0.147)
	0.191[c]	2.294	2.316	2.280
R's education	−0.127	−0.016	−0.010	−0.016
	(0.186)	(0.104)	(0.109)	(0.104)
	−0.680	−0.156	−0.091	−0.159
Father's education	0.008	0.003	0.004	0.003
	(0.010)	(0.005)	(0.006)	(0.005)
	0.871	0.496	0.629	0.489
Father's SES	−0.005	−0.051	−0.033	−0.052
	(0.075)	(0.042)	(0.044)	(0.042)
	−0.066	−1.234	−0.752	−1.245
Collective bargaining = 1	0.356	0.584**	0.428†	0.589**
	(0.470)	(0.262)	(0.274)	(0.261)
	0.758	2.234	1.562	2.264
Urban residence = 1	0.561	0.483**	0.388†	0.485**
	(0.414)	(0.231)	(0.242)	(0.230)
	1.355	2.097	1.607	2.114
Job in the manufacturing sector = 1	0.989**	0.999***	1.041***	0.997***
	(0.437)	(0.244)	(0.255)	(0.243)
	2.262	4.112	4.082	4.124
Married = 1	2.153***	1.221***	1.081***	1.224***
	(0.555)	(0.309)	(0.332)	(0.361)
	3.877	3.953	3.254	3.885
Intercept	4.326**	3.818***	3.560***	3.824***
	(2.056)	(1.145)	(1.200)	(1.141)
	2.104	3.340	2.965	3.358

[a] Parameter estimates.

[b] Standard errors in parentheses (s_n and A_n for the OLS and Bell estimators, respectively).

[c] t-values (Student t on $n-1$ *df* and t^* on the $0.7(n-1)$ *df* for the OLS and Bell estimators, respectively).

†Significant at the .15 level.
*Significant at the .10 level.
**Significant at the .05 level.
***Significant at the .01 level.

Table 14: OLS and Bell estimates for determinants of untransformed wages for 19 year old white males employed and out of school, full ($n = 120$) sample.[a]

	OLS	Bell
Ability score	−6.621	0.341**
	(4.815)	(0.160)
	−1.375	2.145
R's education	−2.605	−0.012
	(3.418)	(0.114)
	−0.762	−0.102
Father's education	0.113	0.003
	(0.176)	(0.006)
	0.638	0.430
Father's SES	0.586	−0.045
	(1.374)	(0.046)
	0.427	−0.982
Collective bargaining = 1	−1.055	0.730**
	(8.612)	(0.287)
	0.591	2.571
Urban residence = 1	4.486	0.568**
	(7.596)	(0.252)
	0.591	2.266
Job in the manufacturing sector = 1	−0.068	1.097***
	(8.021)	(0.267)
	−0.009	4.166
Married = 1	25.511**	1.538***
	(10.192)	(0.339)
	2.503	4.605
Intercept	20.519	3.693***
	(37.720)	(1.254)
	0.588	2.974

[a] See notes to Table 13.

Table A1: OLS, Bounded Influence (BIF), and Bell estimates for determinants of mortality rates in U.S. metropolitan areas ($n = 60$).[a]

	OLS	BIF	Bell
Percent Nonwhite	3.35***	2.60***	2.72***
	(0.59)	(0.67)	(0.48)
	5.68	3.88	5.75
Mean years of	−13.28*	−13.67**	−13.84**
education	(6.98)	(6.12)	(5.68)
	−1.90	−2.23	−2.42
Population (1000's)	2.82	7.13†	4.49†
per square mile	(3.76)	(4.68)	(3.07)
	0.75	1.52	1.47
Precipitation	1.64***	2.01***	1.90***
	(0.62)	(0.44)	(0.50)
	2.66	4.57	3.83
Log SO_2	13.80***	13.61***	15.36***
	(3.82)	(4.10)	(3.11)
	3.61	3.32	5.00
Intercept	930.09***	915.23***	922.17***
	(96.23)	(30.97)	(78.43)
	9.66	11.30	12.12

[a] Parameter estimates, standard errors in parentheses (s_n for the OLS and BIF estimators and A_n for Bell estimator), and t-values (Student t on $n - 1$ *df* for the OLS and BIF estimators and t^* on $0.7(n - 1)$ *df* for Bell estimator).

†Significant at the .15 level.
*Significant at the .10 level.
**Significant at the .05 level.
***Significant at the .01 level.

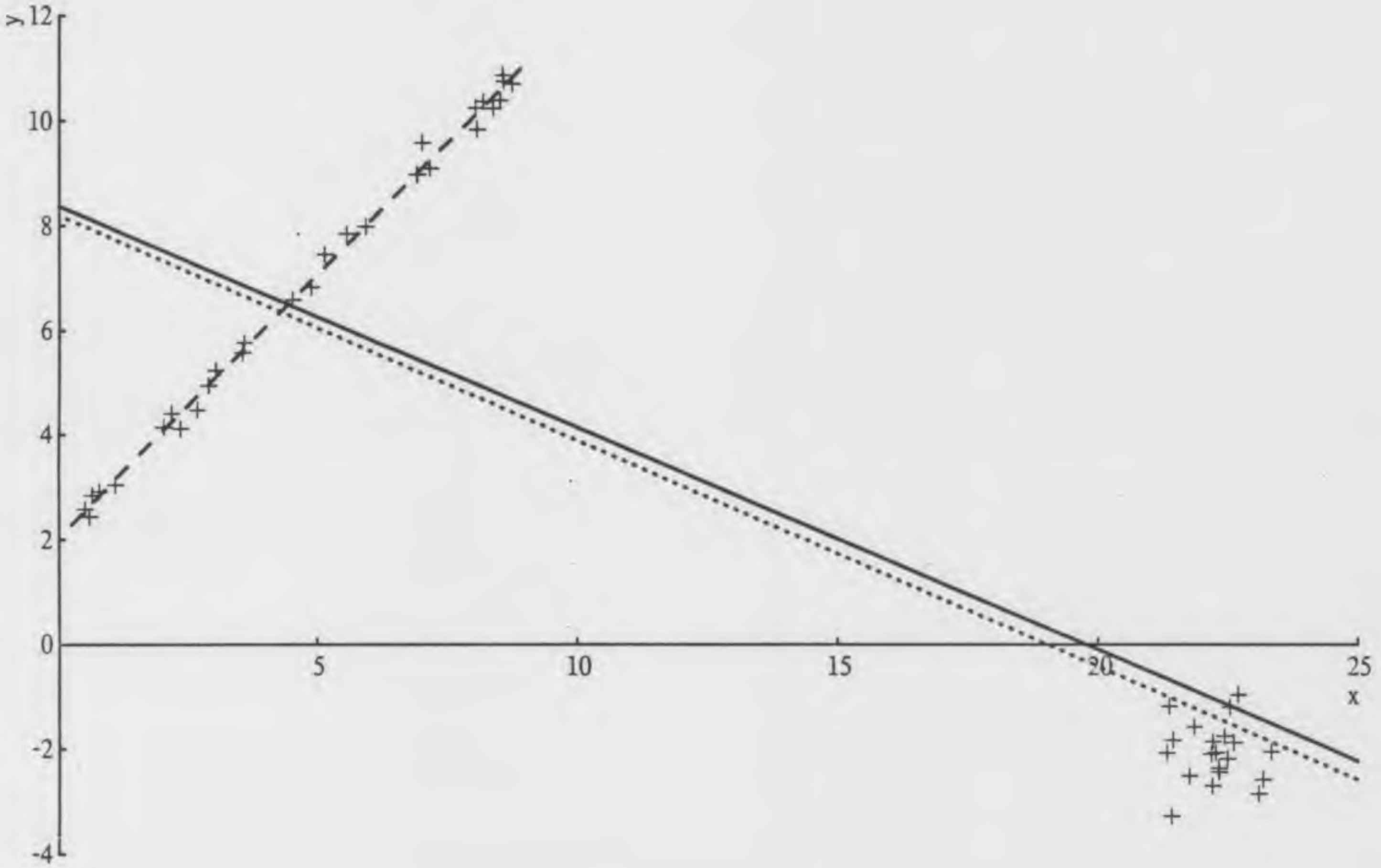


Figure 1

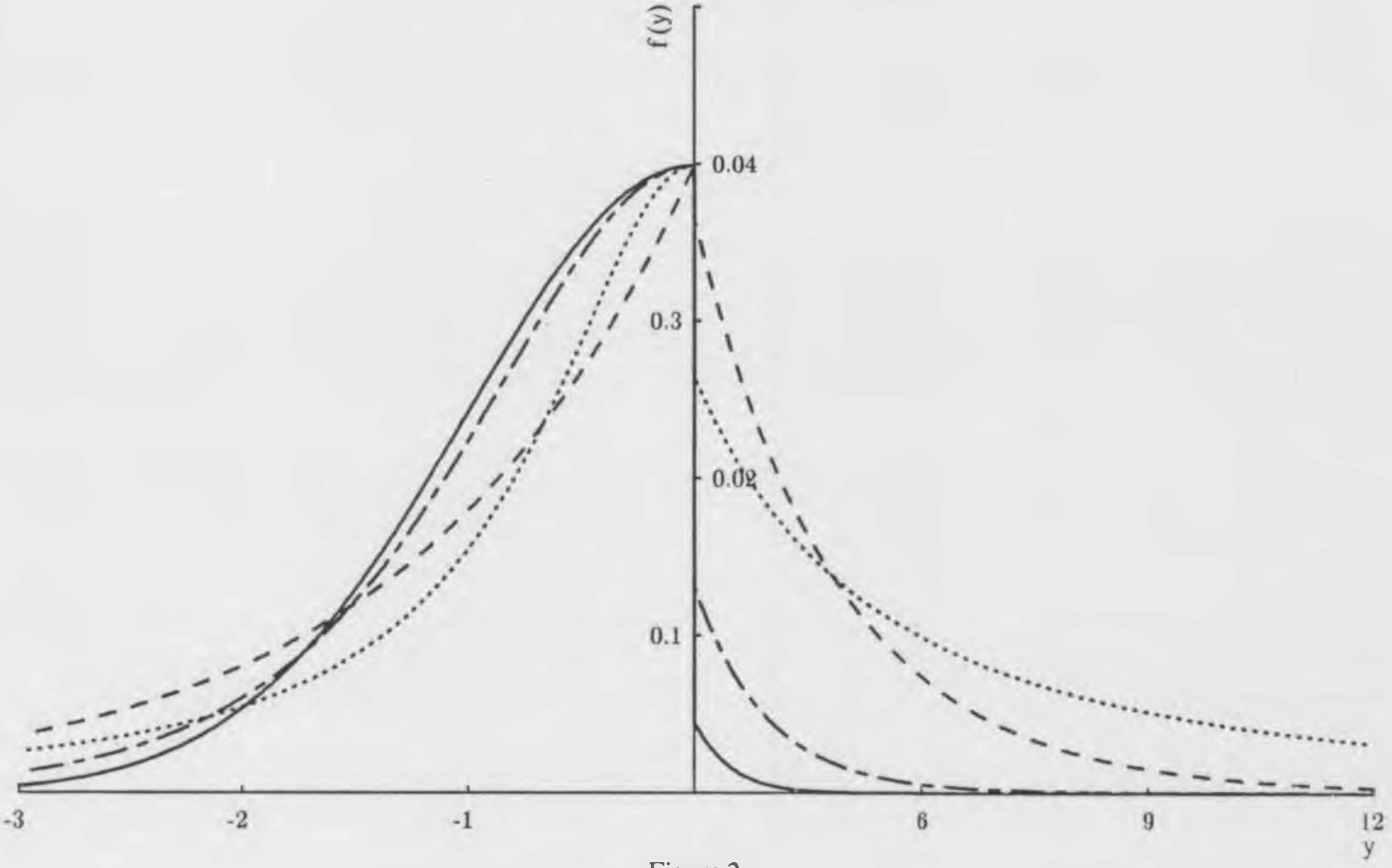


Figure 2

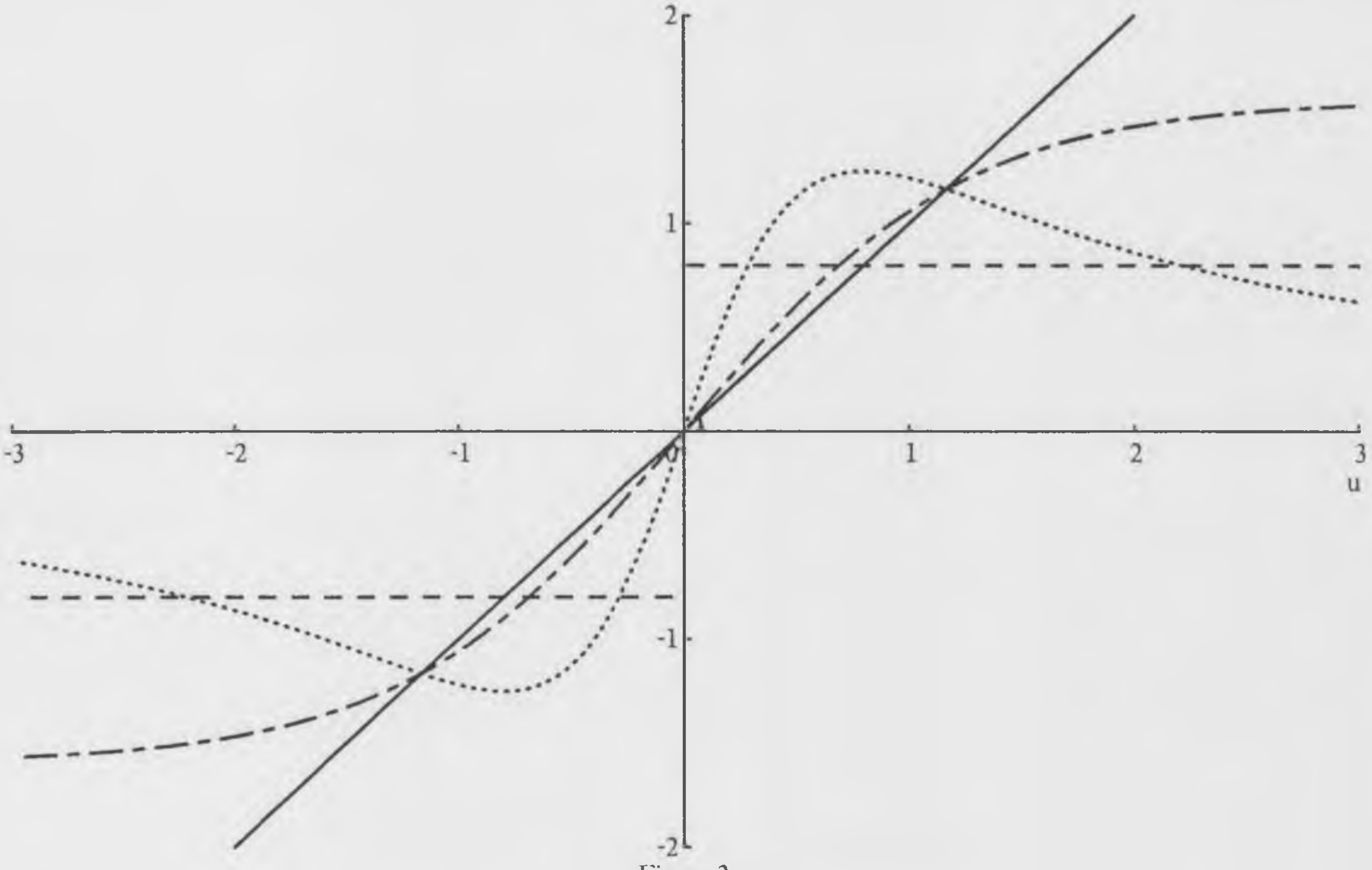


Figure 3

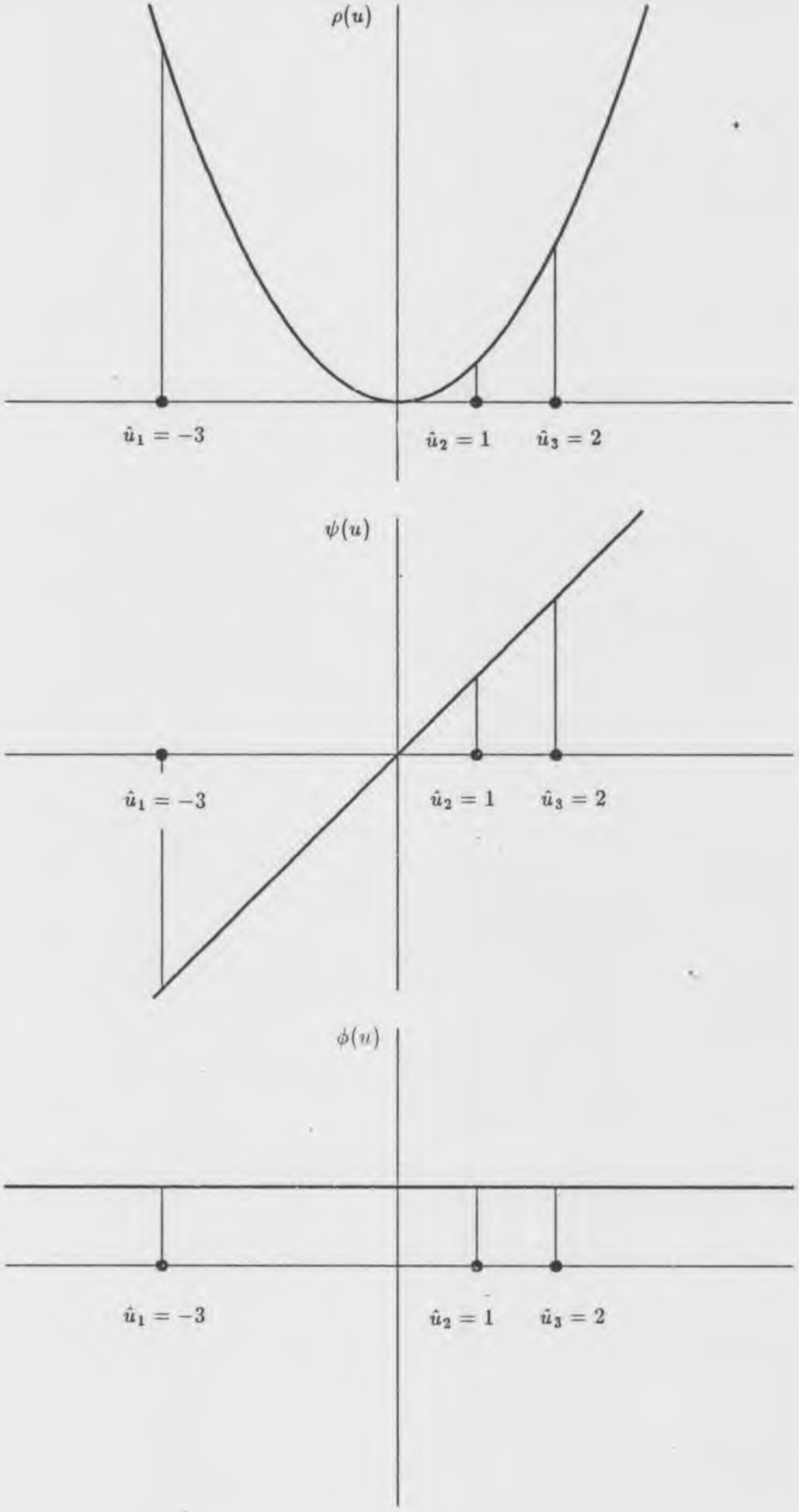

$\rho(u)$
$\hat{u}_1 = -3$
$\hat{u}_2 = 1$
$\hat{u}_3 = 2$
$\psi(u)$
$\hat{u}_1 = -3$
$\hat{u}_2 = 1$
$\hat{u}_3 = 2$
$\phi(u)$
$\hat{u}_1 = -3$
$\hat{u}_2 = 1$
$\hat{u}_3 = 2$

Figure 4

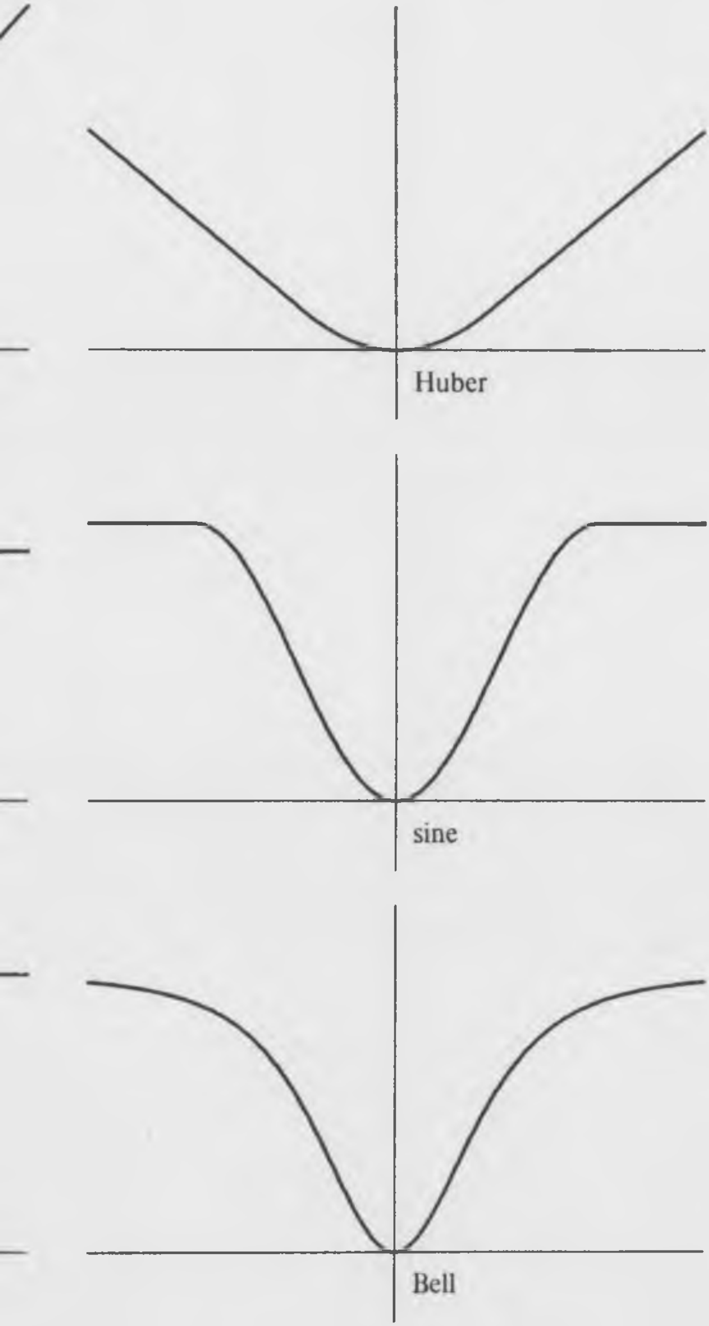
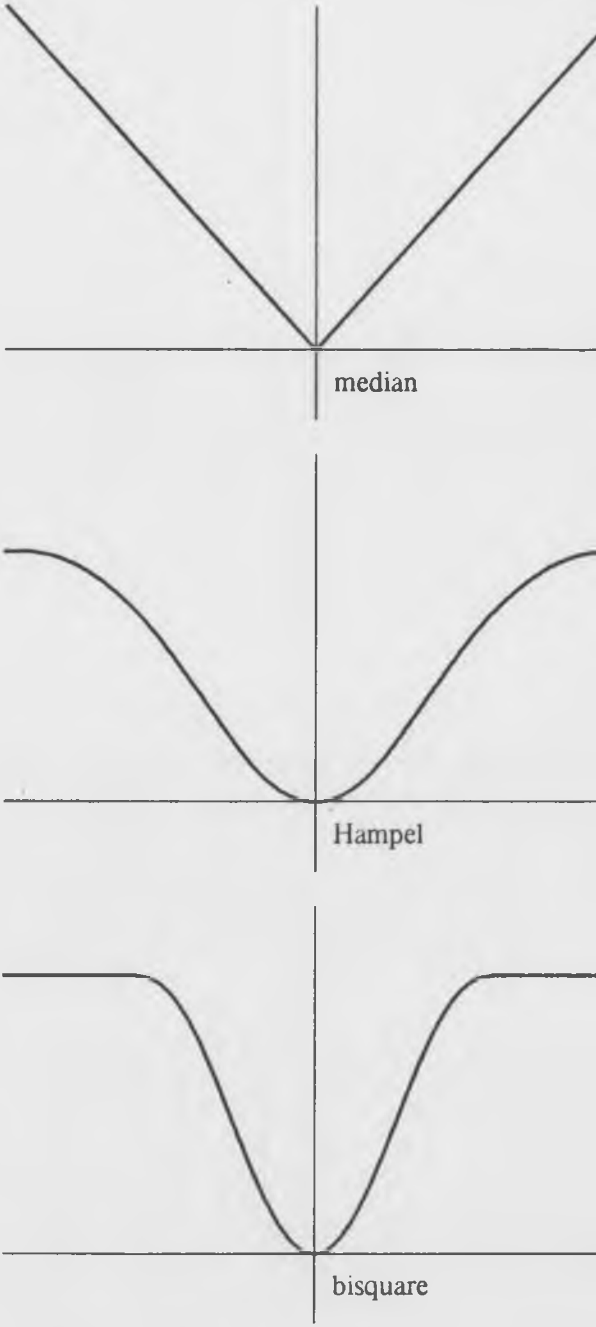


Figure 5


median
Hampel
bisquare

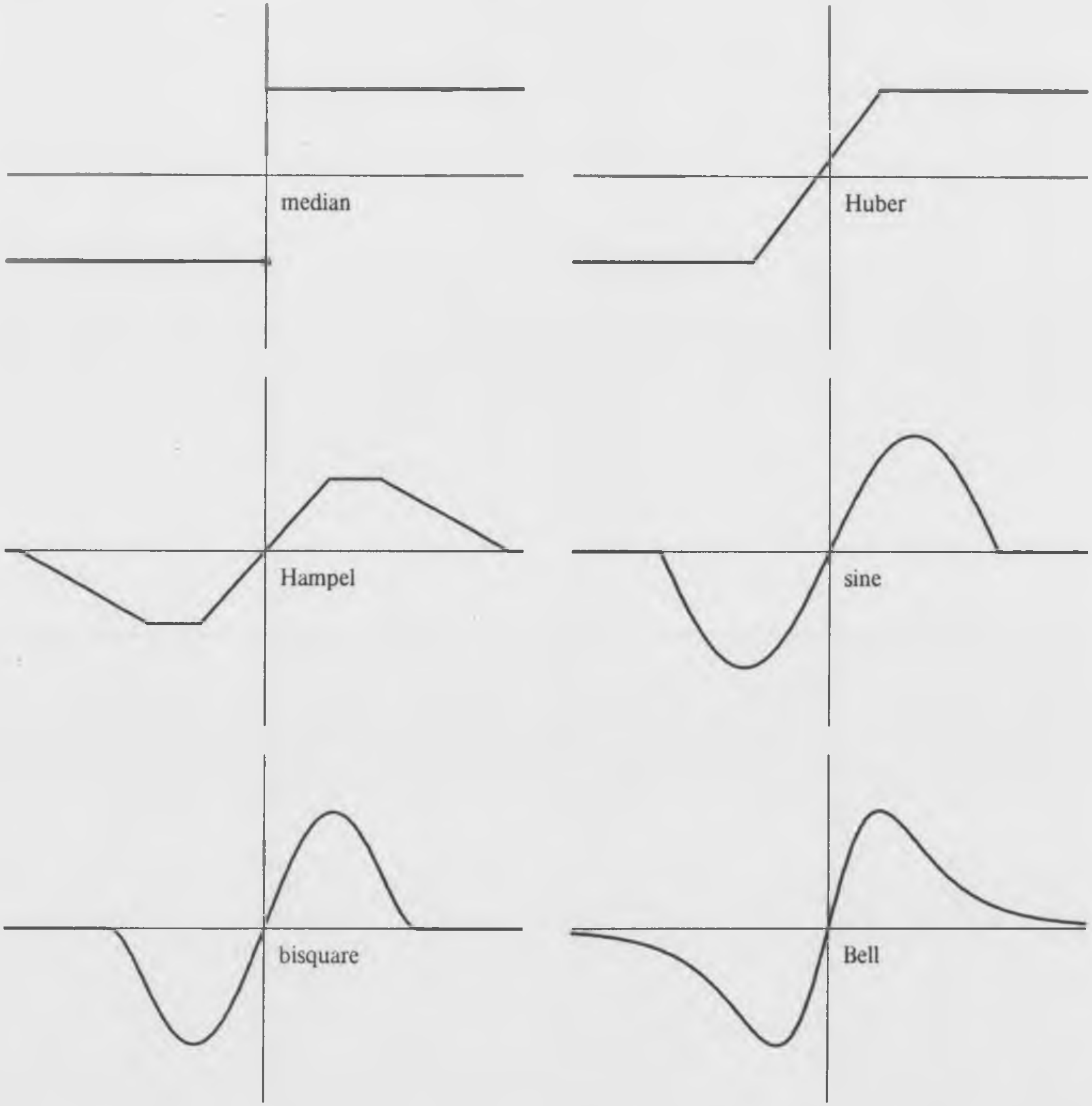
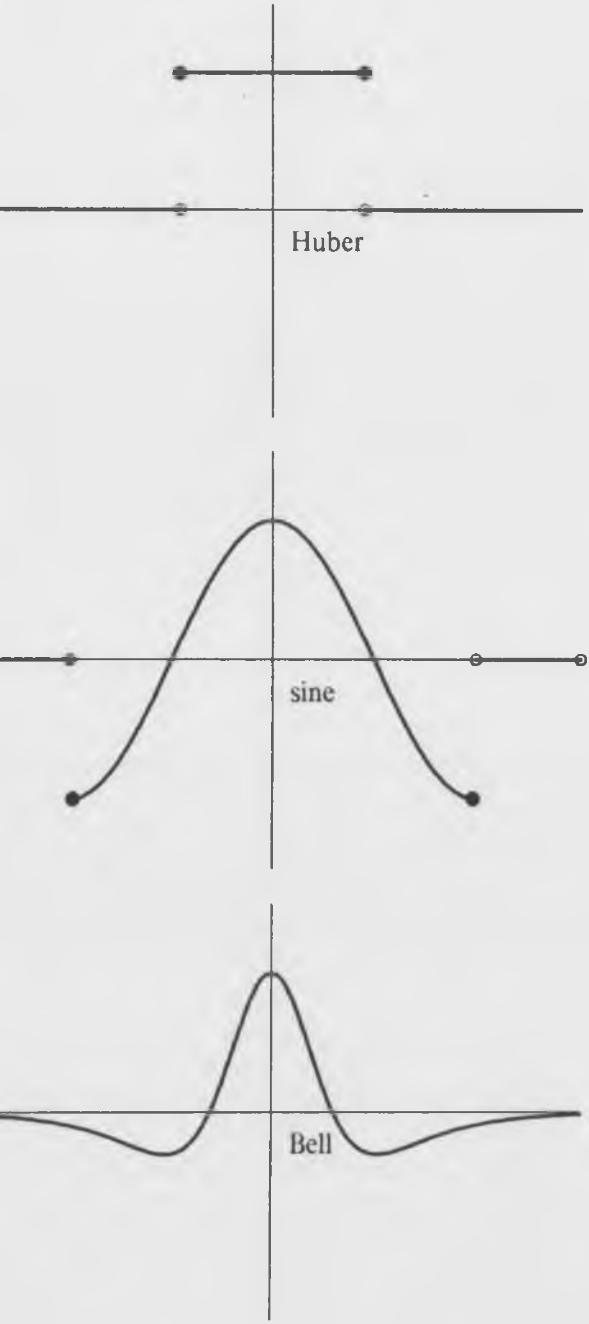


Figure 6


Huber
sine
Bell

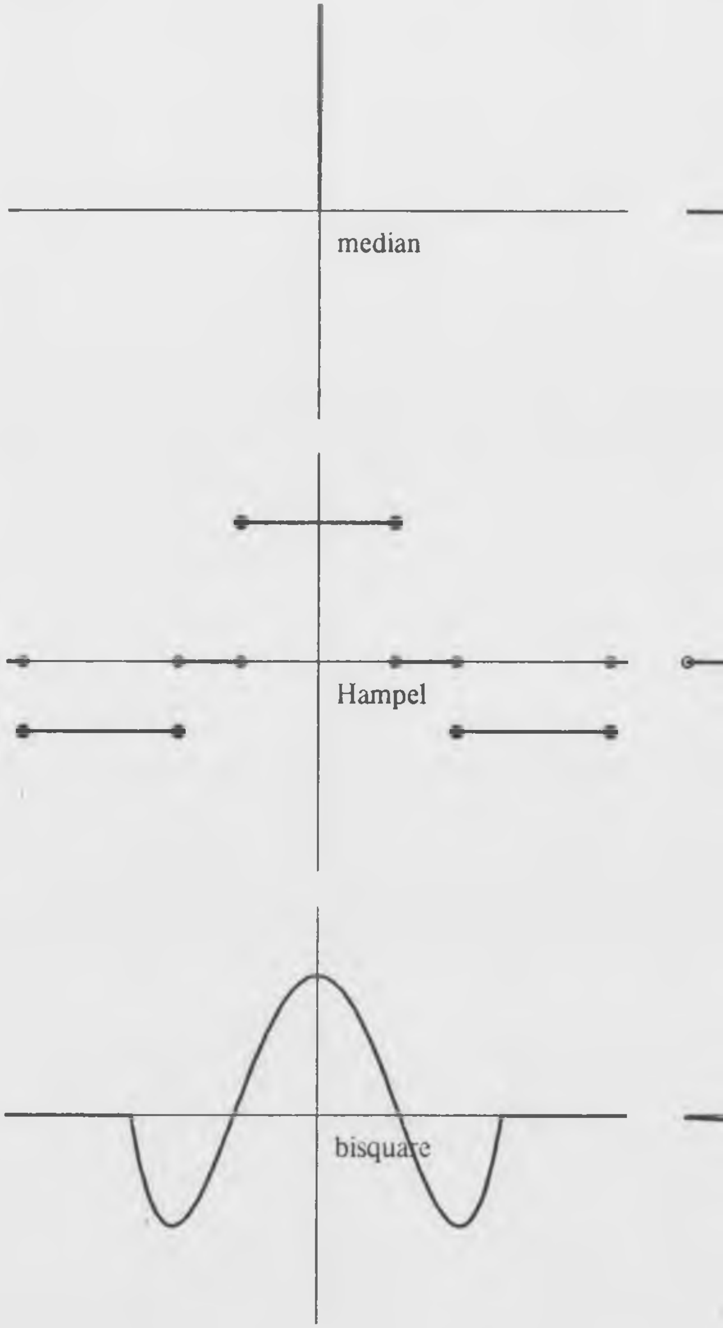


Figure 7

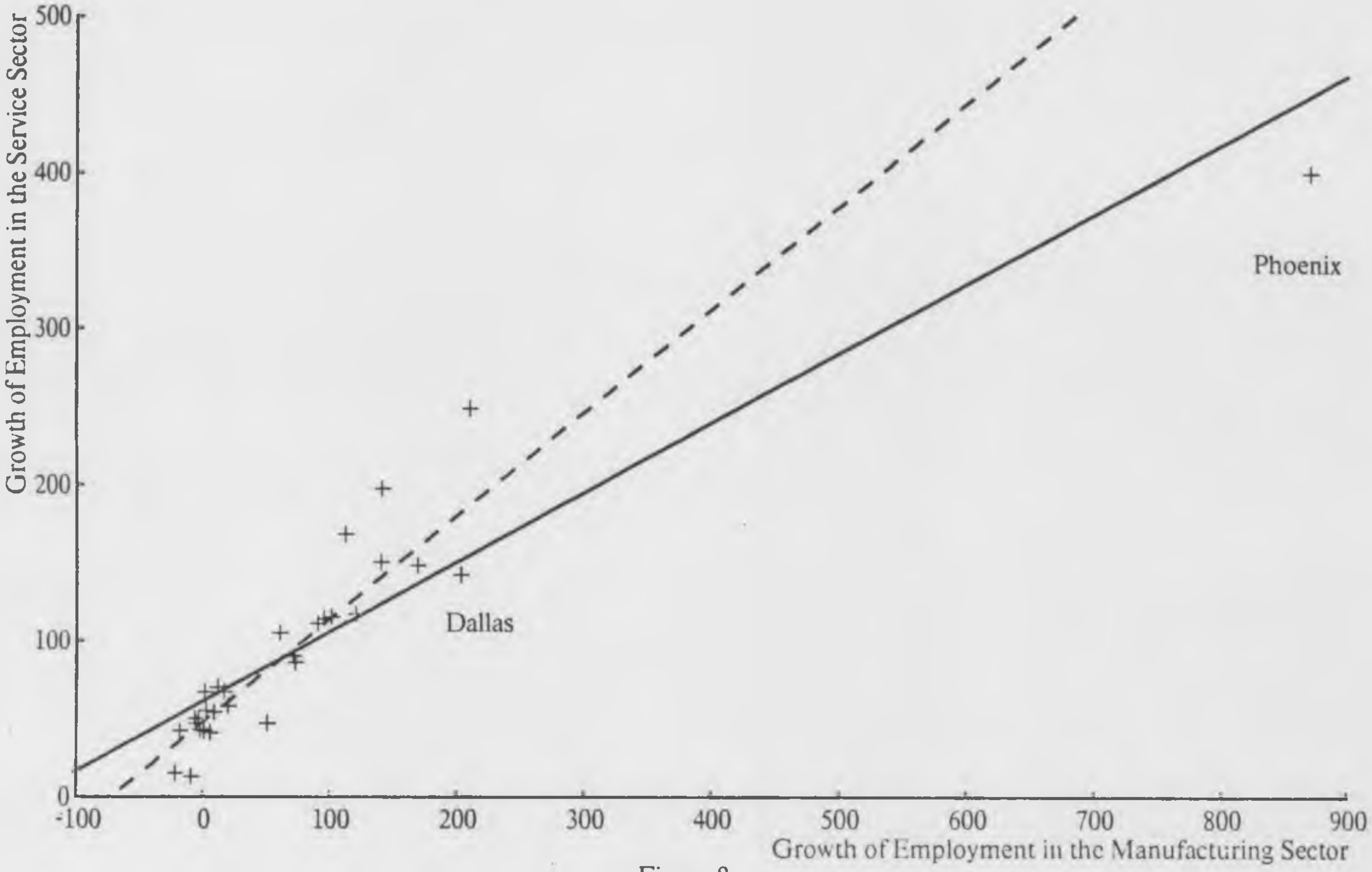
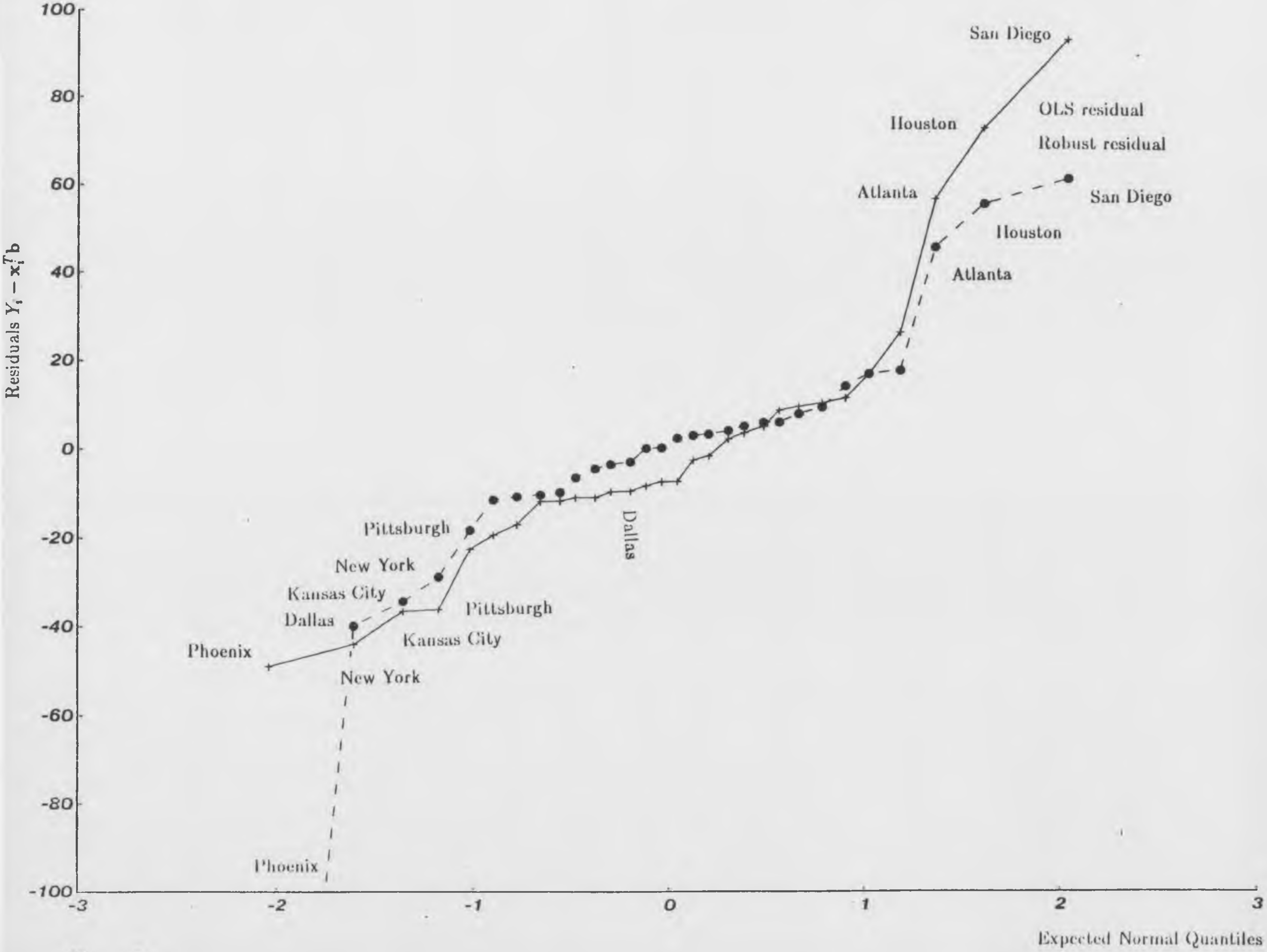

500
400
300
200
100
0
Growth of Employment in the Service Sector
-100
0
100
200
300
400
500
600
700
800
900
Growth of Employment in the Manufacturing Sector
Phoenix
Dallas

Figure 8


San Diego
OLS residual
Robust residual
Houston
Atlanta
San Diego
Houston
Atlanta
Pittsburgh
Dallas
New York
Kansas City
Dallas
Pittsburgh
Kansas City
Phoenix
New York
Phoenix
100
80
60
40
20
0
-20
-40
-60
-80
-100
Residuals $Y_i - \mathbf{x}_i^T\mathbf{b}$
-3
-2
-1
0
1
2
3
Expected Normal Quantiles

(*n* = 120)

Depths	Hourly	wages (units = 1979 dollars)
1	1.t	5
1	1.·	7
3	2.*	0 2 2
6	.t	3 3 5 5 5 5
8	2.·	7 8 8 9 9 9 9 9
16	3.*	0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 3
14	.t	3 3 4 4 4 4 5 5 5 5 5 5 5 5
15	3.·	6 7 7 7 7 7 7 8 8 8 8 8 9 9 9
10	4.*	0 0 0 0 1 2 2 2 2 2
8	.t	3 4 4 5 5 5 5 6
4	4.·	6 7 7 7
5	5.*	0 0 0 0 2
7	.t	3 5 5 5 5 5 6
3	5.·	7 8 9
5	6.*	0 0 0 1 2
5	.t	3 3 4 5 5
3	6.·	6 8 9
0	7.*	
2	.t	5 5
1	7.·	7
1	8.*	1
1	.t	3
1	outlier	(404)

Figure 10

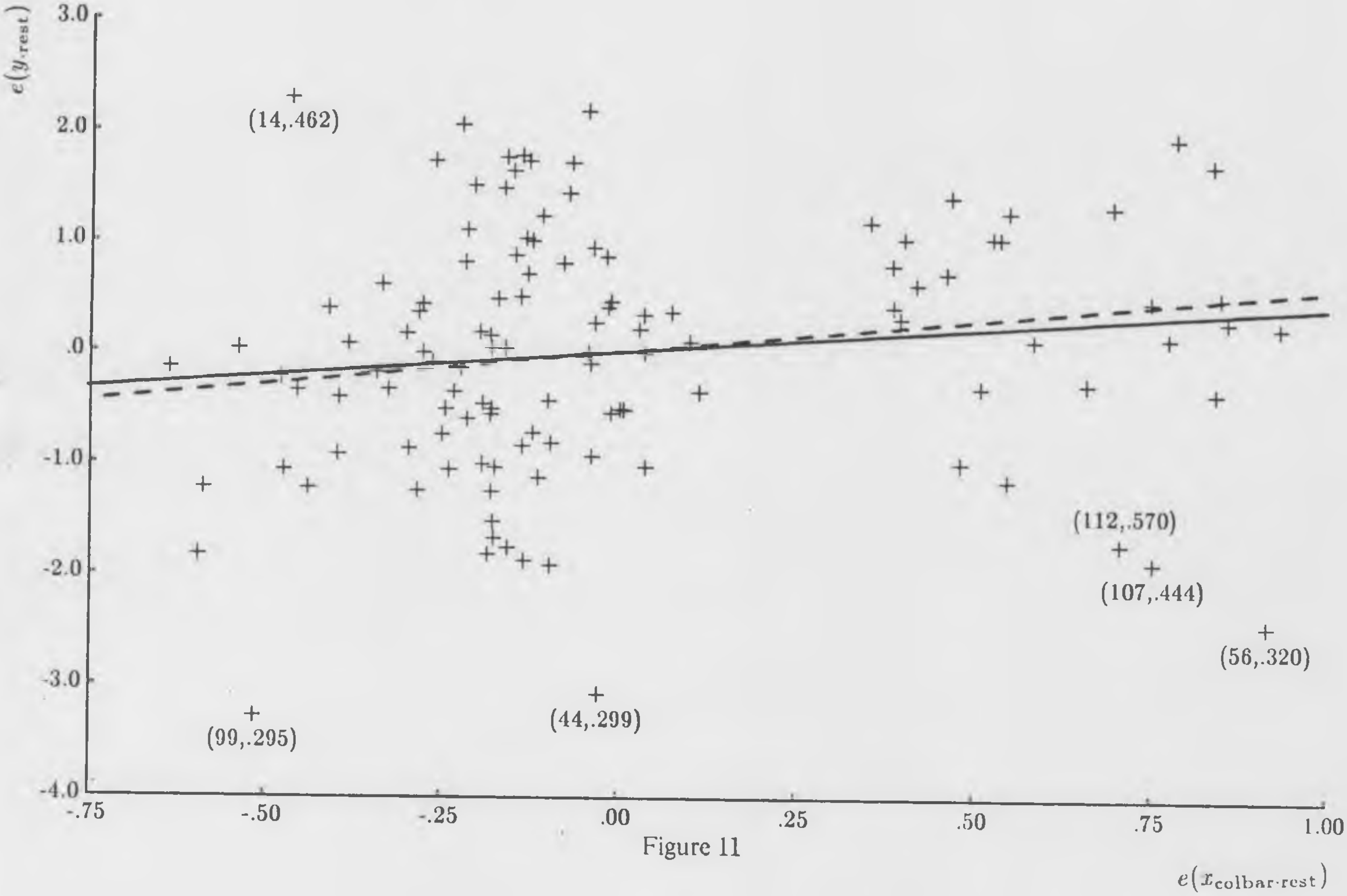

$e(y_{\cdot\text{rest}})$
3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
-.75
-.50
-.25
.00
.25
.50
.75
1.00
$e(x_{\text{colbar}\cdot\text{rest}})$
(14,.462)
(112,.570)
(107,.444)
(56,.320)
(99,.295)
(44,.299)

Figure 11

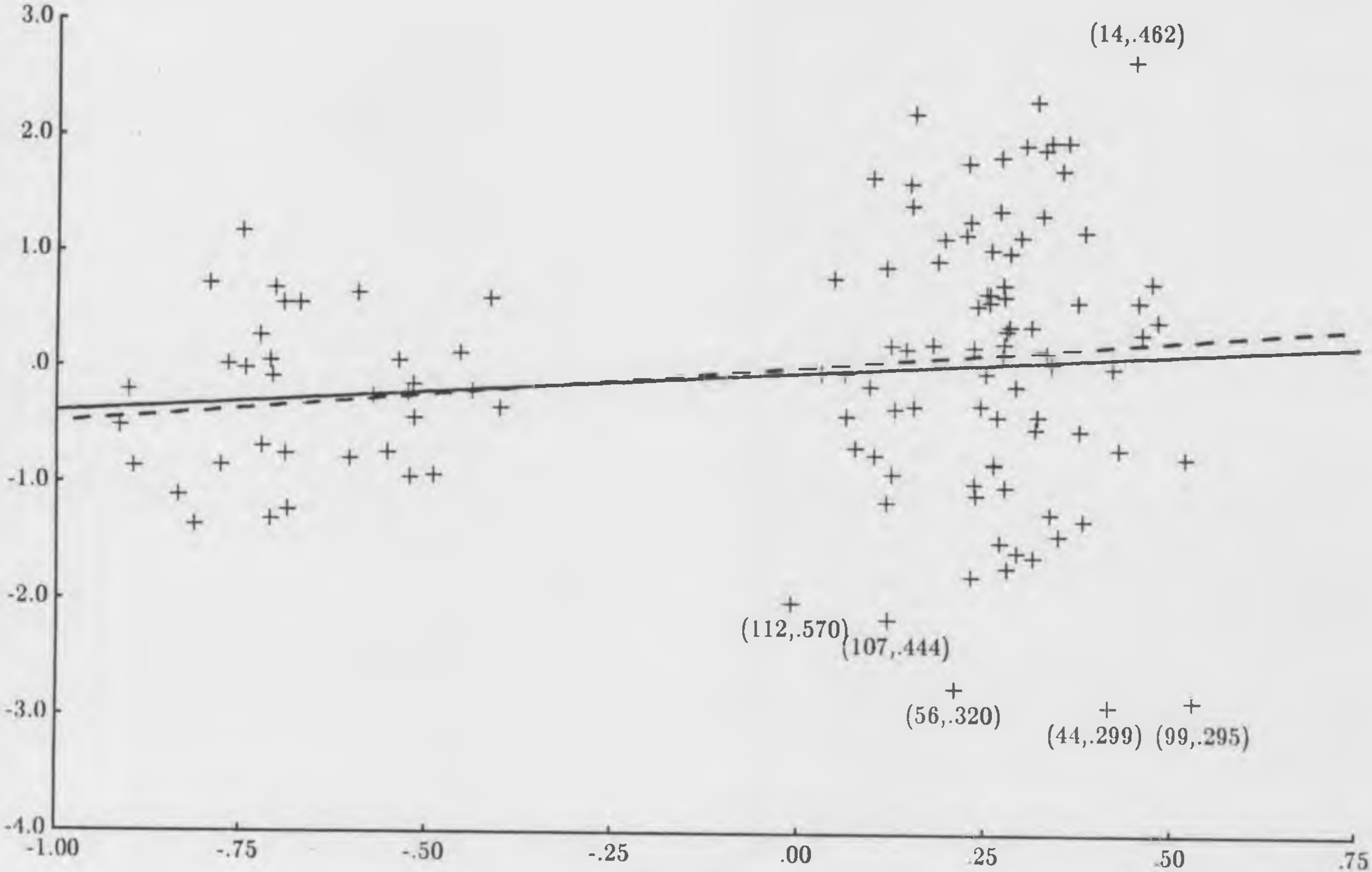

3.0
2.0
1.0
.0
-1.0
-2.0
-3.0
-4.0
-1.00
-.75
-.50
-.25
.00
.25
.50
.75
(14,.462)
(112,.570)
(107,.444)
(56,.320)
(44,.299)
(99,.295)

Figure 12