STATISTICAL ROBUSTNESS: ONE VIEW OF ITS USE IN APPLICATIONS TODAY*

Robert V. Hogg

ABSTRACT

Users of statistical packages should be more aware of the influence that outlying data points can have on their statistical analyses. Robust procedures provide formal methods to spot these outliers and reduce their influence. While a few robust procedures are mentioned in this article, one is emphasized; and it is motivated by maximum likelihood estimation to make it seem more natural. Use of it in regression problems is considered in some detail and an approximate error structure is stated for the robust estimates of the regression coefficients. A few examples are given. A suggestion of how these techniques should be implemented in practice is included.

KEY WORDS

M-estimator, Outliers, Robustness, Robust Regression.

*Reprinted with permission of The American Statistician.

Introduction

Certainly the method of least squares and generalizations of it have served us well for many years. However, it is recognized that "outliers," which arise from heavy tailed distributions or are simply bad data points due to errors, have an unusually large influence on the least squares estimators. That is, the outliers pull the least squares "fit" towards them too much, and a resulting examination of the residuals is misleading because then they look more like normal ones. Accordingly, robust methods have been created to modify least squares procedures so that the outliers have much less influence on the final estimates.

Since George Box (1953) coined the term "robustness," an enormous amount of material has been published on the subject. Certainly John Tukey's (1962) comments on spotty data in his "The future of data analysis" have spurred on these investigations. Perhaps, the greatest contributions have been those of Peter Huber. Clearly his fundamental paper (1964) and later his Wald lectures, and the review articles (1972,1973) based on these talks, are milestones in this development. Also Frank Hampel's use of the influence curve in robust estimation is a central concept, but one beyond the scope of this article. Hence the reader interested in influence curves is referred to Hampel's 1974 paper.

Thus today's statisticians can find plenty of material on M-, R-, and L-estimators and various generalizations, including adaptive versions of them. One survey of some of these procedures is that of Hogg (1974). In addition, a new paperback by Huber (1977) provides an excellent summary of many of the mathematical aspects of robustness.

So, with all of these existing robust methods, we might think that the applications would be flooded with their use. But clearly this is not the case! Of course, some use of them has been made in practice, but most persons continue to use only least squares techniques (or generalizations of least squares) that are associated with usual normal assumptions. Yet if one studies the literature on robustness, it does seem as if there is some place for these newer techniques. Seemingly, today we should seriously question the dogma of normality in each analysis under consideration and truly be concerned about the influence of outliers and bad data points on our inferential statistics. This concern should range from simple cases with data that is not clearly out-of-line to multivariate situations in which it is extremely difficult to detect spotty data. Thus we need formal procedures—like robust methods—to help find those outlying points and eliminate or reduce their effects.

Exactly why is it that robust methods are not used more today? Possibly there are too many robust schemes, and the applied statistician simply does not know where to start.

Thus it seems to me that the "robustniks" should agree to support certain basic and understandable robust procedures and then try to sell them to the statistical community. This does not mean that all research in this exciting area should be stopped. On the contrary, it should be encouraged and stimulated, with the understanding that recommendations made in future years will quite likely be different from those of today. But we should try to make a significant start in the introduction of robust estimation in statistical practice now; so the major purpose in writing this article is to generate some interest in robustness among statistical users.

Thus this article is <u>not</u> a survey of the many robust techniques. As a matter of fact, we will consider a very limited number of procedures and concentrate on only one type—incidentally one that, with some adaptations, can be used whenever "least squares" (or generalizations of it) is used: regression, ANOVA, multivariate analysis, discrimination, etc. This proposed method is a very reasonable approach with fairly easy computations; thus I believe it should be the major one in use today.

In the article, I will try to provide some background, beginning with maximum likelihood estimation of one parameter, that will hopefully make robust estimation seem natural and rather easy. Some approximate sampling distribution theory will be stated for these robust estimators; thus something will be known about the error structure so that statistical inferences can be made. Finally, I must emphasize here that I am not going to recommend that we discontinue using the method of least squares and all of the computer packages associated with it and its generalizations. On the contrary, I strongly urge that we continue to use these methods with which we are so familiar. But I also urge that, along with each such analysis, a robust one should also be made. Then if the two procedures are in essential agreement, report that fact and the usual (least squares) summary of the analysis. But if substantial differences exist in the two analyses, another hard look at the data must be taken, searching in particular for outliers or bad data points. And the robust procedures will, for all practical purposes, detect these spotty data points for you by recording low weights for large residuals from the robust fit.

The Location Parameter

In a distribution described by the density $f(x)$, let us introduce an unknown location (slippage) parameter θ, obtaining the density $f(x-\theta)$, $-\infty < \theta < \infty$. If $X_1, X_2, \cdots, X_n$ represents a random sample from this distribution, one popular method of estimating θ is that of maximum likelihood. The logarithm of the likelihood function $L(\theta)$ is

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i-\theta) = -\sum_{i=1}^{n} \rho(x_i-\theta),$$

where $\rho(x) = -\ln f(x)$. If we can maximize by differentiating, we have

$$\frac{d \ln L(\theta)}{d\theta} = -\sum_{i=1}^{n} \frac{f'(x_i-\theta)}{f(x_i-\theta)} = \sum_{i=1}^{n} \psi(x_i-\theta),$$

where $\rho'(x) = \psi(x)$. The solution of

$$\sum_{i=1}^{n} \psi(x_i-\theta) = 0$$

that maximizes $L(\theta)$ is called the maximum likelihood estimator of θ and is frequently denoted by $\hat{\theta}$.

Three typical classroom examples of this process are given by the following distributions, the first of which provides the least squares estimate and the second, the "least absolute values" estimate.

(1) <u>Normal</u>: $\rho(x) = \frac{x^2}{2} + c$, $\psi(x) = x$,

$$\sum_{i=1}^{n} (x_i-\theta) = 0 \quad \text{yields} \quad \hat{\theta} = \bar{x}.$$

(2) <u>Double exponential</u>: $\rho(x) = |x| + c$, $\psi(x) = \begin{cases} -1, & x < 0, \\ 1, & x > 0, \end{cases}$

$$\sum_{i=1}^{n} \psi(x_i-\theta) = 0 \quad \text{yields} \quad \hat{\theta} = \text{sample median}.$$

(3) Cauchy: $\rho(x) = \ln(1+x^2) + c$, $\psi(x) = \dfrac{2x}{1+x^2}$,

$\sum_{i=1}^{n} \psi(x_i-\theta) = 0$ is solved by iterative methods.

It is interesting to note that the ψ functions of examples (2) and (3) are bounded and that of (3) even redescends and approaches zero asymptotically. We note that the solutions in (2) and (3) are not influenced much by outliers. On the other hand, the least squares estimator $\overline{X}$ of (1) is greatly influenced by extreme values. That is, it is well known that the least squares estimator $\overline{X}$ is not extremely good in situations in which the underlying distribution has long tails, for example, in the Cauchy case. Therefore, in robust estimation, we look for estimators that are quite efficient (usually around 90 to 95 percent) if the underlying distribution is normal, but are also very efficient even though the underlying distribution has long tails. Sometimes the amount of efficiency lost under normal assumptions, say 5%, is referred to as the premium which we pay for the protection that we get in case we actually have a distribution that has longer tails than the normal one.

For a certain theoretical reason (actually minimizing the maximum asymptotic variance of the estimators associated with a certain class of distributions), Huber (1964) proposed that we use for our robust estimator the maximum likelihood estimator of the location parameter associated with a density that is like a normal in the middle but like a double exponential in the tails. In particular, Huber's ρ function is (except for an additive constant)

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 , & |x| \le k, \\ k|x| - \frac{1}{2}k^2 , & k < |x|, \end{cases}$$

so that the ψ function is

$$\psi(x) = \begin{cases} -k , & x < -k, \\ x , & -k \le x \le k, \\ k , & k < x. \end{cases}$$

Of course, with this ψ function, the equation

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0$$

must be solved by iterative methods. An estimator of this type (not necessarily using this particular ψ function) is denoted by $\hat{\theta}$ and called an M-estimator, M for maximum likelihood.

There is another feature of this M-estimator on which we should comment. Suppose a solution $\hat{\theta}$ was found for a particular sample $x_1, x_2, \cdots, x_n$, and then these items were replaced by some in which the deviations from $\hat{\theta}$ were tripled, for example. The new solution $\hat{\theta}$, using this modified sample, would not necessarily be the same. That is, the estimator is not scale invariant. To obtain a scale invariant version of this estimator, we could solve

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \theta}{d}\right) = 0,$$

where now d is a robust estimate of scale. While ad hoc, a popular statistic d used in this solution is

$$d = \text{median}|x_i - \text{median}(x_i)|/(0.6745).$$

(Sometimes the numerator of d is called the MAD, the median of the absolute deviations.) The divisor 0.6745 is used because then $d \approx \sigma$ if n is large and if the sample actually arises from a normal distribution. Usually the sample standard deviation s is not used as a d value since it is influenced too much by outliers and thus is not robust.

This particular scheme of selecting d suggests appropriate values of the "tuning" constant k so that the efficiency of $\hat{\theta}$ will be high if the underlying distribution is actually normal. In the normal situation, we would want most of the items to satisfy the inequality

$$\left|\frac{x_i - \theta}{d}\right| \leq k$$

because then

$$\psi\left(\frac{x_i-\theta}{d}\right) = \frac{x_i-\theta}{d} .$$

As a matter of fact, if all items enjoyed this inequality, then $\hat{\theta} = \overline{X}$ which is the desired estimator in the normal case. Since $d \approx \sigma$, k is usually taken to be some number around 1.5. When $k = 1.5$, we refer to this procedure (or the corresponding estimator) as a (1.5)Huber procedure (estimator). If σ is known (that is, d is known), the asymptotic efficiency of it, under normal assumptions, is greater than 95% (Huber, 1964); and, in most heavy tailed situations, it performs extremely well (Andrews et al., 1972). Thus our premium is small for much protection in nonnormal cases. However, σ is usually unknown and it must be estimated; this does reduce that efficiency only a very little when $n \geq 10$; see Andrews et al. (1972).

Other ψ functions that are commonly used are the following, along with suggested values of the respective tuning constants.

(a,b,c) Hampel

$$\psi(x) = (\text{sign } x)\begin{cases} |x| & , \quad 0 \leq |x| < a, \\ a & , \quad a \leq |x| < b, \\ a\frac{c-|x|}{c-b} & , \quad b \leq |x| < c, \\ 0 & , \quad c \leq |x|. \end{cases}$$

Reasonably good values of the constants are $\underline{a} = 1.7$, $\underline{b} = 3.4$, $\underline{c} = 8.5$.

(k)Wave of Andrews

$$\psi(x) = \begin{cases} \sin(x/k) , & |x| \le k\pi \\ 0 , & |x| > k\pi \end{cases}$$

with $\underline{k} = 1.5$ or 2.0. Actually if the scale is known, $\underline{k} = 1.339$ requires a premium of 5%.

(k)Biweight of Tukey

$$\psi(x) = \begin{cases} x[1-(x/k)^2]^2 , & |x| \le k, \\ 0 , & |x| > k, \end{cases}$$

with $\underline{k} = 5.0$ or 6.0. If the scale is known, $\underline{k} = 4.685$ implies a premium of 5%.

It should be noted that the Wave and Biweight procedures are very similar and are reasonable substitutes for each other. Since the ρ functions associated with these three redescending ψ functions is not convex, there could be certain convergence problems in the iterative procedures, although this is not too likely. That is, these have been successful procedures and should be used, but with some care.

Suppose, by an iterative numerical method, we find the solution $\hat{\theta}$ to our equation; that is, let $\hat{\theta}$ be the M-estimator so that $\Sigma\psi[(X_i-\hat{\theta})/d] = 0$, for any bounded ψ function which is odd. It is easy to approximate this last equation by replacing the left-hand member by two terms of

Taylor's series, expanded about the true parameter value θ. From this approximation, when $d = \sigma$, it is a straightforward exercise to show

$$\sqrt{n}(\hat{\theta}-\theta) \approx \frac{\sigma\left[\Sigma\psi\left(\frac{X_i-\theta}{\sigma}\right)\right]\Big/\sqrt{n}}{\left[\Sigma\psi'\left(\frac{X_i-\theta}{\sigma}\right)\right]\Big/n}$$

and this expression has a limiting normal distribution with mean zero and variance

$$\frac{\sigma^2 E\left[\psi^2\left(\frac{X-\theta}{\sigma}\right)\right]}{\left\{E\left[\psi'\left(\frac{X-\theta}{\sigma}\right)\right]\right\}^2}\,.$$

Of course, since we do not know the underlying density, we must, in practice, approximate the expected values and the σ that appear in this asymptotic variance. One approximation to the variance of $\sqrt{n}(\hat{\theta}-\theta)$ is given by

$$s_1^2 = \frac{d^2\left[\frac{1}{n}\sum_{i=1}^{n}\psi^2\left(\frac{x_i-\hat{\theta}}{d}\right)\right]}{\left[\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{x_i-\hat{\theta}}{d}\right)\right]^2}\,.$$

That is, for large n, $\sqrt{n}(\hat{\theta}-\theta)/s_1$ has an approximate standardized normal distribution. An even better approximating distribution to $\sqrt{n-1}(\hat{\theta}-\theta)/s_1$ can be found by using one member of the t-family, possibly one with $n-1$ degrees of freedom (or even one with somewhat smaller degrees of freedom than n-1); see Huber (1970). With one of these approximate distributions, we can make statistical inferences about the unknown θ. Gross (1976) explores confidence intervals based on this idea as well as those resulting from other schemes, and he finds robustness of validity (maintaining the 95% confidence coefficient) and of efficiency (establishing relatively short intervals).

Robust Regression

Suppose that we have the linear model:

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{E}$$

where $\underset{\sim}{Y}$ is a nx1 random vector, $\underset{\sim}{X}$ is a nxp design matrix of known constants such that $\underset{\sim}{X}'\underset{\sim}{X}$ is of full rank, $\underset{\sim}{\beta}$ is a px1 vector of unknown parameters, and $\underset{\sim}{E}$ is a nx1 random vector, whose elements are like a random sample from a distribution that is symmetric (but not necessarily normal) about zero. To parallel the approach used with the location parameter, we wish to minimize, with a robust ρ function, the summation

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}}{d}\right)$$

where y_i is the i^{th} element of $\underset{\sim}{Y}$, $\underset{\sim}{x}_i$ is the i^{th} row of $\underset{\sim}{X}$ and d is an estimate of the scale of the distribution associated with $\underset{\sim}{E}$. Equating the first partial derivatives with respect to the elements of $\underset{\sim}{\beta}$, say β_j, equal to zero, we see that this is equivalent to finding the maximizing solution associated with the p equations

$$\sum_{i=1}^{n} x_{ij}\,\psi\left(\frac{y_i - \underset{\sim}{x}_i\,\underset{\sim}{\beta}}{d}\right) = 0, \quad j = 1,2,\cdots,p,$$

where x_{ij} is the element in the i^{th} row and j^{th} column of $\underset{\sim}{X}$.

Let us first concern ourselves with an initial estimate $\underset{\sim}{\beta}_0$ of $\underset{\sim}{\beta}$; this is needed for two things: (a) a robust estimate d of scale and (b) a "start" in the iteration to find $\hat{\underset{\sim}{\beta}}$. If it is easy to find the L_1 ("least absolute values") estimators for the regression coefficients, these would be good and something like the median in the single sample case. Moreover, for the estimate of scale, we would then use the median of the absolute values of the nonzero residuals divided by 0.6745, as in the location case. Many statisticians, however, find it inconvenient to determine L_1 estimators, and therefore let us consider an algorithm of Dutter (1977) for estimating $\underset{\sim}{\beta}$ and scale simultaneously. It is described here only for the ψ function of Huber. This method, as will be seen, is then very close to least squares (and is only one of several suggestions listed in Dutter's article).

Let us begin with some initial estimates, $\underset{\sim}{\beta}_0$ and d_0, which might very well be the usual estimates of $\underset{\sim}{\beta}$ and scale. We proceed as follows:

1. Compute the residuals $z_i = y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}_0$, $i = 1,2,\cdots,n$.

2. Find a new estimate of scale

$$d_1^2 = \frac{1}{(n-p)E(\psi^2)} \sum_{i=1}^{n} \left[\psi\left(\frac{z_i}{d_0}\right)\right]^2 d_0^2,$$

   where $E(\psi^2)$ is the expected value of Huber's $\psi^2(W)$, where W is a standardized normal random variable. For illustration, $E(\psi^2) = 0.7785$ in the (1.5)Huber procedure.

3. "Winsorize" the residuals; that is, determine

$$\Delta_i = \psi\left(\frac{z_i}{d_1}\right) d_1 , \quad i = 1,2,\cdots,n.$$

   Of course, the Winsorized residual $\psi(z_i/d_1)d_1$ equals z_i if $|z_i/d_1| \le k$ but is equal to kd_1 $(-kd_1)$ if $kd_1 < z_i$ $(z_i < -kd_1)$. For more on the general concept of Winsorizing, see Dixon and Tukey (1968).

4. Find the least squares estimates of the regression coefficients as if the Winsorized residuals were the observations, namely

$$\underset{\sim}{\hat{\tau}}_0 = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\Delta}_0$$

   where $\underset{\sim}{\Delta}_0$ is a px1 column vector of $\Delta_1,\Delta_2,\cdots,\Delta_n$.

5. Compute a new estimate of $\underset{\sim}{\beta}$, namely

$$\underset{\sim}{\beta}_1 = \underset{\sim}{\beta}_0 + q\underset{\sim}{\hat{\tau}}_0 .$$

   Dutter found that a suitable choice of the factor q

is $q = \min\left[\frac{1}{\Phi(k)-\Phi(-k)}, 1.9\right]$, where Φ is the standardized normal distribution function.

With the new estimates, $\underset{\sim}{\beta}_1$ and d_1, as the starting values, repeat steps 1-5. Continue this iteration process until (on the m^{th} iteration), for all $i = 1,2,\cdots,p$,

$$|q\hat{\tau}_m^i| < \epsilon d_{m+1}\sqrt{x_{ii}} \quad \text{and} \quad |d_{m+1}-d_m| < \epsilon d_{m+1},$$

where $\epsilon > 0$ is an appropriate tolerance level, $\hat{\tau}_m^i$ is the i^{th} component of $\underset{\sim}{\hat{\tau}}_m$, and x_{ii} is the i^{th} diagonal element of $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$. At that point, stop the iterations and estimate $\underset{\sim}{\beta}$ and the scale parameter using $\underset{\sim}{\beta}_{m+1}$ and d_{m+1}, respectively.

The reader should note exactly how close this procedure is to least squares. Of course, step (4) is least squares on the Winsorized residuals and step (5) would provide the least square estimate of $\underset{\sim}{\beta}$ in case $q = 1$ and $\underset{\sim}{\Delta}_0$ equals the vector of actual residuals (not Winsorized) for any initial estimate $\underset{\sim}{\beta}_0$. It is also of interest to observe that $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ needs to be computed only once in the iterative process and this is a definite computational saving.

After obtaining good robust estimates of $\underset{\sim}{\beta}$ and scale using Dutter's algorithm, least absolute values, or another scheme (some nonparametric methods might be quite suitable), we could treat outliers more severely using these robust estimates as new "starts," say $\underset{\sim}{\tilde{\beta}}_0$ and $\tilde{\sigma}_0$, with a redescending ψ function such as a Wave or Biweight. In those cases, a "weighted least squares" procedure is a good algorithm to use. In this method we replace the p equations

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}}{d}\right) = 0, \quad j = 1,2,\cdots,p,$$

or, equivalently when $y_i - \underset{\sim}{x}_i\underset{\sim}{\beta} \neq 0$,

$$\sum_{i=1}^{n} x_{ij}\left[\frac{\psi\left(\frac{y_i - \underset{\sim}{x}_i\underset{\sim}{\beta}}{d}\right)}{\frac{y_i - \underset{\sim}{x}_i\underset{\sim}{\beta}}{d}}\right](y_i - \underset{\sim}{x}_i\underset{\sim}{\beta}) = 0,$$

by the approximations

$$\sum_{i=1}^{n} x_{ij}(w_{i0})(y_i - \underset{\sim}{x}_i\underset{\sim}{\beta}) \approx 0, \quad j = 1,2,\cdots,p,$$

where

$$w_{i0} = \frac{\psi\left(\frac{y_i - \underset{\sim}{x}_i\underset{\sim}{\tilde{\beta}}_0}{\tilde{d}_0}\right)}{\frac{y_i - \underset{\sim}{x}_i\underset{\sim}{\tilde{\beta}}_0}{\tilde{d}_0}}, \quad y_i \neq \underset{\sim}{x}_i\underset{\sim}{\tilde{\beta}}_0$$

and $w_{i0} = 1$, when $y_i = \underset{\sim}{x}_i\underset{\sim}{\tilde{\beta}}_0$. In matrix notation, in which $\underset{\sim}{W}_0$ is the nxn diagonal matrix with $w_{10}, w_{20}, \cdots, w_{n0}$ on the principal diagonal, the one-step estimator is

$$\underset{\sim}{\tilde{\beta}}_{1W} = (\underset{\sim}{X}'\underset{\sim}{W}_0\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{W}_0\underset{\sim}{Y}.$$

The iteration requires that, on each step, we recompute the weights and thus the inverse $(\underset{\sim}{X}'\underset{\sim}{W}_j\underset{\sim}{X})^{-1}$, $j = 0,1,2,\cdots$. However, with good estimates $\underset{\sim}{\tilde{\beta}}_0$ and $\tilde{d}_0$ resulting from Dutter's procedure, only a few iterations are usually needed to obtain good redescending ψ estimates.

What should be used for the error structure associated with the final estimate $\underset{\sim}{\hat{\beta}}$? It is true (Huber, 1973) that under certain reasonable conditions (one is a known spread $d = \sigma$), $\underset{\sim}{\hat{\beta}}$ has an approximate normal distribution with mean $\underset{\sim}{\beta}$ and variance-covariance matrix

$$\frac{\sigma^2 E\left[\psi^2\left(\frac{Z}{\sigma}\right)\right](\underset{\sim}{X}'\underset{\sim}{X})^{-1}}{\left\{E\left[\psi'\left(\frac{Z}{\sigma}\right)\right]\right\}^2},$$

where Z represents an element of the random vector $\underset{\sim}{E}$. An approximation to this variance-covariance matrix is

$$\frac{(nd^2)\left[\frac{1}{n}\Sigma\,\psi^2\left(\frac{y_i - \underset{\sim}{x}_i\underset{\sim}{\hat{\beta}}}{d}\right)\right](\underset{\sim}{X}'\underset{\sim}{X})^{-1}}{(n-p)\left[\frac{1}{n}\Sigma\,\psi'\left(\frac{y_i - x_i\hat{\beta}}{d}\right)\right]^2}.$$

Of course, the weighted least squares program also automatically provides another estimate of this variance-covariance matrix. Two other suggestions are given by Welsch (1975), but there is not general agreement on which of these approximations is best. Nevertheless, whichever one we choose, we do have some idea about the error structure of $\underset{\sim}{\hat{\beta}}$, and we can thus make some approximate statistical inferences about $\underset{\sim}{\beta}$ using the usual normal theory.

Let us close this section on regression with a remark about the procedure when we do not have a linear model, but

$$\underset{\sim}{Y} = h(\underset{\sim}{\beta}) + \underset{\sim}{E},$$

where $h(\cdot)$ is a nonlinear function of $\underset{\sim}{\beta}$. Let $h_i(\underset{\sim}{\beta})$ be $h(\underset{\sim}{\beta})$ which is associated with Y_i. We wish to minimize

$$\sum_{i=1}^{n} \rho\left(\frac{y_i-h_i(\underset{\sim}{\beta})}{d}\right)$$

where an estimate d of spread can be found using a preliminary estimate $\tilde{\underset{\sim}{\beta}}_0$. Equating the first partial derivatives to zero, we have

$$\sum_{i=1}^{n} \frac{\partial h_i(\underset{\sim}{\beta})}{\partial \beta_j}\, \psi\left(\frac{y_i-h_i(\underset{\sim}{\beta})}{d}\right) = 0, \quad j = 1,2,\cdots,p.$$

With the preliminary estimate $\tilde{\underset{\sim}{\beta}}_0$ and a weighted nonlinear least squares package, it is easy to solve

$$\sum_{i=1}^{n} \frac{\partial h_i(\underset{\sim}{\beta})}{\partial \beta_j}\, (w_{i0})[y_i-h_i(\underset{\sim}{\beta})] = 0, \quad j = 1,2,\cdots,p,$$

where

$$w_{i0} = \frac{\psi\left(\dfrac{y_i-h_i(\tilde{\underset{\sim}{\beta}}_0)}{d}\right)}{\dfrac{y_i-h_i(\tilde{\underset{\sim}{\beta}}_0)}{d}}, \quad y_i \neq h_i(\tilde{\underset{\sim}{\beta}}_0)$$

and $w_{i0} = 1$, when $y_i = h_i(\tilde{\underset{\sim}{\beta}}_0)$, $i = 1,2,\cdots,n$. In the nonlinear case, the "weighted least squares" algorithm seems to be easiest although Dutter's algorithm can be modified with $\partial h_i/\partial \beta_j$ replacing x_{ij}.

Holland and Welsch (1977) compare robust estimates resulting from eight ψ functions. They also note that a semiportable subroutine library (including these eight ψ functions) called ROSEPACK (RObust Statistical Estimation PACKage) has been developed at the Computer Research Center of the National Bureau of Economic Research, Inc. in Cambridge, Mass. For example, from their report, a "tuning constant" of $k = 1.5$ seems to be appropriate for the Wave estimate because then it would have an efficiency greater than 95%, provided $d = \sigma$ is known and provided the underlying distribution is normal. One desirable feature of a redescending ψ function, like that of Andrews' Wave or Tukey's Biweight, is that the extreme outliers are treated much more harshly than in Huber's procedure; and many times, in practice, we find that zero weight has been assigned to an outlier or bad data point.

Of course, whenever least squares procedures (or generalizations of it) are used, It seemingly would be possible to use robust procedures with some type of adaptation. These include ANOVA, regression, time series, splines, multivariate analysis, and discrimination. For example, Lenth (1977) has produced some robust splines that give excellent fits to data points for which the usual (least squares) splines fail. Also Randles et al. (1977) have found robust estimates of mean vectors and variance-covariance matrices that are used in discrimination problems. Robust ANOVA procedures are reported on by

Schrader and McKean (1977). From these studies it is clear that, with the necessary imagination, appropriate adaptations can be made; and, hopefully, much more will be done in the future.

L-Estimation

Let us first consider the simple case of a random sample from a distribution of the continuous type that has a location parameter θ. Say the order statistics of the sample are $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. An L-estimator is one which is a linear combination of these order statistics. Examples of L-estimators are:

(a) sample median,

(b) α-trimmed mean $\overline{X}_\alpha = \frac{1}{n-2[n\alpha]} \sum_i X_{(i)}$, where the summation is over $i = [n\alpha]+1, \cdots, n-[n\alpha]$,

(c) Gastwirth's estimator which is a weighted average of the $33\frac{1}{3}$rd, 50^{th}, and $66\frac{2}{3}$rd percentiles with respective weights .3, .4, and .3,

(d) Tukey's trimean which is a weighted average of the 1^{st}, 2^{nd}, and 3^{rd} quartiles with respective weights $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$.

These and other L-estimators are also described by Andrews et al. (1972).

The generalization of L-estimators to the regression situation is not clear as in the case of M-estimators. However, since the use of the ρ function, $\rho(x) = |x|$, yields the median as an estimator (and the "median plane or surface" in the regression situation), this could easily be modified to get other percentiles. That is, the ρ function

$$\rho(x) = \begin{cases} -(1-p)x, & x < 0, \\ px \quad , & x \geq 0, \end{cases}$$

yields the $(100p)^{th}$ percentile in the single sample case and thus estimates of the "$(100p)^{th}$ percentile plane or surface" in the regression situation; see Koenker and Bassett (1978). Clearly, generalizations of estimates like those of Gastwirth and Tukey could now be constructed in regression problems. Moreover, it seems as if in many situations (for example, educational data involving prediction of college performance from high school rank and SAT or ACT scores) we would be interested in estimates of some percentiles other than those of the middle. Thus percentile estimates could stand on their own as well as in combination with others to predict a "middle" plane or surface.

R-Estimation

R-estimation is a nonparametric method resulting from ranking when the sample arises from a continuous-type distribution. It can easily be extended to regression. Consider the linear model and modify least squares by replacing one factor in $(y_i - \underset{\sim}{x}_i \underset{\sim}{\beta})^2$ by the rank of $y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}$, say R_i. The rank R_i is clearly a function of $\underset{\sim}{\beta}$. Hence we wish to

$$\text{minimize } \sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}) R_i .$$

This, in turn, can be generalized by replacing the ranks $1,2,\cdots,n$ by the "scores"

$$a(1) \le a(2) \le \cdots \le a(n).$$

Thus, in this generalized setting, we wish to

$$\text{minimize } \sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i \underset{\sim}{\beta}) a(R_i).$$

Of course, two examples of scores are

(a) Wilcoxon scores: $a(i) = i$ or ranks,

and

(b) Median scores: $a(i) = \begin{cases} -1, & i < (n+1)/2, \\ 1, & i > (n+1)/2. \end{cases}$

Jaeckel (1972) proved that this minimization is equivalent to solving the p-equations

$$\sum_{i=1}^{n} x_{ij} a(R_i) = 0, \quad j = 1,2,\cdots,p,$$

that must be solved approximately due to the discontinuities in $a(\cdot)$ and R_i. Moreover, it is well known that "good" (having certain asymptotic properties) scores are those given by

$$a(i) = \varphi\left(\frac{i}{n+1}\right), \quad \text{where} \quad \varphi(t) = -\frac{f'[F^{-1}(t)]}{f[F^{-1}(t)]} .$$

Examples of this are:

(a) f normal produces $\varphi(t) = \Phi^{-1}(t)$, $0 < t < 1$, that gives normal scores;

(b) f double exponential produces $\varphi(t) = \begin{cases} -1, & 0 < t < \frac{1}{2} \\ 1, & \frac{1}{2} < t < 1, \end{cases}$ that gives median scores;

(c) f logistic produces $\varphi(t) = 2t-1$, $0 < t < 1$, that gives Wilcoxon scores.

Jurečková (1977) proved that, with certain scores $a(\cdot)$ and ψ functions, the R-estimators and M-estimators are asymptotically equivalent. Among other conditions we need that

$$\varphi(t) = c_1\psi[F^{-1}(t)] + c_2,$$

where c_1 and c_2 are constants, for this equivalence. In light of this result, it seems more reasonable to use M-estimators as the computations are easier, at least at this time.

Adaptive Estimators

Only brief note is made here of adaptive estimators; but, if interested in more background, the reader is referred to an expository article on the subject by Hogg (1974). The basic idea of adapting is the selection of the estimation procedure after observing the data. Thus, for example, the tuning constants or the amounts of trimming could be dictated by the sample. As a matter of fact, even the forms of the function $\psi(\cdot)$ and the score function $a(\cdot)$ could be selected after observing the sample. Of course, asymptotically, we can select the "best" $\psi(\cdot)$ or $a(\cdot)$, but most of the time we are working with sample sizes like 20, 30, or 50, not infinity. Hence we must find some reasonable procedures for those very limited sample sizes. One such scheme is to select a small class of underlying distributions that span a large collection of possible distributions. Then determine a good procedure for each member of that class. Finally let the observations select (through analysis of residuals, plots, etc.) which procedure will actually be used by taking that distribution seemingly closest to what was observed. Incidentally, there is no objection to analyzing with all of the procedures: If they say the same thing, that is the answer; but if they differ, then the selection procedure is critical and it is most important that it be done well.

There has been some evidence (Hogg, 1974) that adaptive procedures are of value. After all, if the "Hubers, Hampels, Waves, Biweights,..." are good, wouldn't adaptive ones be better (particularly since the former are included in the latter)? Moreover, it seems as if the applied statisticians would find adaptation very appealing (they do it all the time anyway), and this gives us a chance to bring theory and applications closer together.

Examples

1. Linear Regression. Andrews (1974) reports on a set of data that had been analyzed by Daniel and Wood (1971). There were 21 observations and 3 independent variables. After some astute observations, Daniel and Wood were able to set aside 4 of these 21 observations because of unusual behavior. At the end of their analysis, they did use a different model from

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

However, Andrews notes that "most researchers do not have the insight and perseverance of these authors," Hence he simply summarizes three fits to this model: (A) Least squares on all 21 points, (B) Least squares on 17 points after discarding those 4 points, (C) Robust estimates using the Andrews' ψ function with $k = 1.5$ on all 21 points (the estimated standard errors are in parentheses).

Method	β_1	β_2	β_3
A	0.72(0.17)	1.30(0.37)	-0.15(0.16)
B	0.80(0.07)	0.58(0.17)	-0.07(0.06)
C	0.82(0.05)	0.52(0.12)	-0.07(0.04)

It is very impressive to note that the robust method on all 21 points provides essentially the same estimates as does least squares using the 17 observations, with the 4 bad points set aside. This means that an investigator could have used least squares (A) and the robust scheme (C) and found the four spotty points and the better estimates without having the "insight and perseverance" of Daniel and Wood.

2. Half-Life of Plutonium-241. Zeigler and Ferris (1973) reported that each of six laboratories had a sample of Plutonium containing ^{238}Pu, ^{239}Pu, ^{240}Pu, ^{241}Pu, and ^{242}Pu. To determine the half-life of ^{241}Pu, the ratio (say Y) of the contents of ^{241}Pu to that of ^{239}Pu was reported by each of six labs every 4 months. This was continued longer

than three years until over 70 data points were collected. The problem was to fit the nonlinear function $E(Y) = \beta_0 e^{-\beta_1 t}$. At a later date and with additional data, Zeigler used, along with least squares, a robust scheme based on Andrews' Wave. The two estimates of half-life were 14.84 and 14.70 years, respectively. However, the most interesting part of this analysis was that, with Andrews' weighting function, there were 6 data points with weights of zero, all six of which were reported incorrectly by the same lab because of some technical error (that has since been corrected).

3. Splines. Lenth (1977) considered 51 observations that were simulated from a Cauchy distribution such that the median of each was on the curve $\sin(2\pi e^{-x^2})$, where the 51 x values ranged from zero to 2.5. The conventional least-squares spline fit with six knots $(x = 0, 0.3, 0.7, 1.2, 1.8, \text{ and } 2.5)$ was compared to two robust spline fits: (1.1)Huber and (1.2)Wave. The latter two produced much better fits than the one by least squares. Of course, the Andrews' Wave fit was somewhat better than that of Huber because a redescending M-estimate is more appropriate with the underlying Cauchy distribution.

4. Automated Data Reduction. Agee and Turner (1978) note that grossly erroneous measurements, when undetected, "completely destroyed automated data reduction" at the U.S. Army

White Sands Missile Range. The application of M-estimation (primarily with Hampel's ψ) has been highly successful in dealing with these problems which occur in data preprocessing, instrument calibration, N-station cinetheodolites, N-station radar situations, and filtering.

Summary

Good applied statisticians have always been on guard for outliers or bad data points, discarding them or investigating them further as is appropriate. However, in complicated data sets, it is most difficult to spot some of these extreme points. But a formal robust procedure can definitely help us in this regard. Hence it is recommended that in our statistical investigations we do the following.

(a) Perform the usual (least squares or a generalization of it) analyses.

(b) Also use a robust procedure. Ideally, this might be (1.5)Huber for several iterations followed by the (1.5)Wave or (5.0)Biweight for two or three iterations. However, at the minimum, the one-step Wave or Biweight estimator (that is, one step of an iterative process) should be found, noting which weights are starting to decrease substantially from the number one.

(c) If estimates from methods (a) and (b) are in essential agreement, report that agreement and the usual statistical summaries associated with (a).

(d) If the estimates from methods (a) and (b) do not agree very well, take another hard look at the data. In particular, look at those points having low weights or large residuals from the robust fit (the weights and residuals of the points should always be displayed on the last iteration). Then the usual questions can be asked about these points: from "Has someone made a simple recording error?" to "Is this outlier trying to tell us something significant about our experiment?"

If a robust element is added to our present day methods, we will detect many simple, and not so simple, errors. These procedures have been used very successfully (for example, Los Alamos Scientific Laboratory has the option of using them in all regression problems and this option is exercised frequently). Many extremely interesting things have been discovered through their use. My hope is that by 1980 almost all statistical investigations will include a robust aspect. And, by that time, the researchers in robust methods will have other new and better procedures to propose to the statistical community.

References

Agee, William S. and Turner, Robert H. (1978), "Application of Robust Statistical Methods to Data Reduction," Technical Report No. 65, White Sands Missile Range, New Mexico.

Andrews, D.F., et al. (1972), Robust Estimates of Location, Princeton, New Jersey: Princeton University Press.

Andrews, D.F. (1974), "A Robust Method for Multiple Linear Regression," Technometrics, 16, 423-531.

Box, G.E.P. (1953), "Non-Normality and Tests on Variances," Biometrika, 40, 318-335.

Daniel, Cuthbert and Wood, Fred S. (1971), Fitting Equations to Data, Wiley-Interscience, New York.

Dixon, W.J. and Tukey, John W. (1968), "Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2)," Technometrics, 10, 83-98.

Dutter, Rudoft (1977), "Numerical Solution of Robust Regression Problems: Computational Aspects, a Comparison," Journal of Statistical Computation and Simulation, 5, 207-238.

Gross, Alan M. (1976), "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," Journal of the American Statistical Association, 71, 409-416.

Hampel, Frank R. (1974), "The Influence Curve and Its Role in Robust Estimation," Journal of the American Statistical Association, 69, 383-393.

Hogg, Robert V. (1974), "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," *Journal of the American Statistical Association*, *69*, 909-927.

Holland, Paul W. and Welsch, Roy E. (1977), "Robust Regression Using Iteratively Reweighted Least-Squares," *Communications in Statistics*, *A6*, 813-828.

Huber, Peter J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, *35*, 73-101.

__________ (1970), "Studentizing Robust Estimates," in M. L. Puri, ed., *Nonparametric Techniques in Statistical Inference*, London, New York: Cambridge University Press, 435-463.

__________ (1972), "Robust Statistics: A Review," *Annals of Mathematical Statistics*, *43*, 1041-1067.

__________ (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, *1*, 799-821.

__________ (1977), *Robust Statistical Procedures*, Society of Industrial and Applied Mathematics, Philadelphia, PA.

Jaeckel, L.A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," *Annals of Mathematical Statistics*, *43*, 1449-1458.

Jurečková, J. (1977), "Asymptotic Relations of M-Estimates and R-estimates in Linear Regression Models," *Annals of Statistics*, 5, 464-472.

Koenker, Roger and Bassett, Gilbert (1978), "Regression Quantities," Econometrics, 46, 33-50.

Lenth, Russell V. (1977), "Robust Splines," Communications in Statistics, A6, 847-854.

Randles, Ronald H., Broffitt, James D., Ramberg, John S., and Hogg, Robert V. (1978), "Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates," Journal of the American Statistical Association, 73, 564-568.

Schrader, Ronald M. and McKean, Joseph W. (1977), "Robust Analysis of Variance," Communications in Statistics, A6, 879-894.

Tukey, John W. (1962), "The Future of Data Analysis," Annals of Mathematical Statistics, 33, 1-67.

Welsch, Roy E. (1975), "Confidence Regions for Robust Regression," Working Paper No. 111, National Bureau of Economic Research, Cambridge, Massachusetts.

Zeigler, R.K. and Ferris, Yvonne (1973), "Half-Life of Plutonium-241," Journal of Inorganic Nuclear Chemistry, 35, 3417-3418.