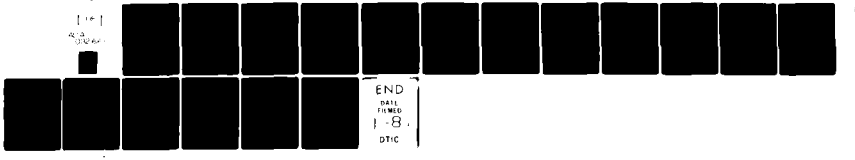
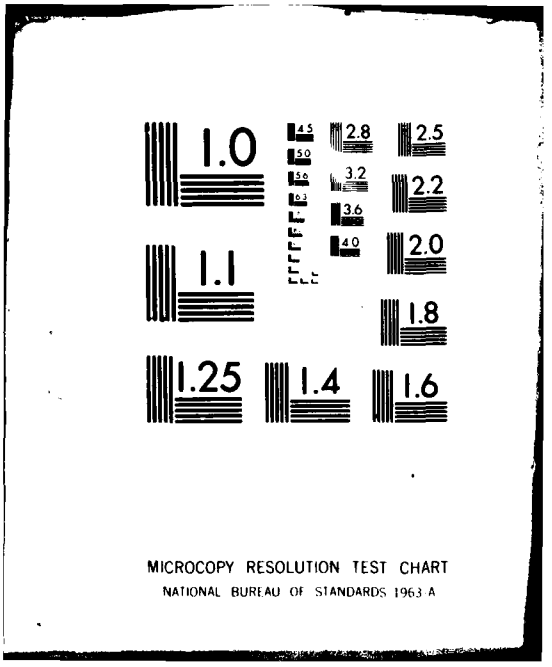
AD-A092 661 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
TEACHING ROBUST METHODS FOR EXPLORATORY DATA ANALYSIS.(U)
OCT 80 A F SIEGEL DAAG29-79-C-0205
UNCLASSIFIED TR-173-SER-2 ARO-16669.3-M NL



END
DATE FILMED
1-81
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 16669.3-M	2. GOVT ACCESSION NO.: AD-A092 661
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Teaching Robust Methods for Exploratory Data Analysis	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Andrew F. Siegel	8. CONTRACT OR GRANT NUMBER(s): DAAG29-79-C-0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Princeton University, Princeton, NJ 08540	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709	12. REPORT DATE: Oct 80 13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

statistics
estimation
teaching
data analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

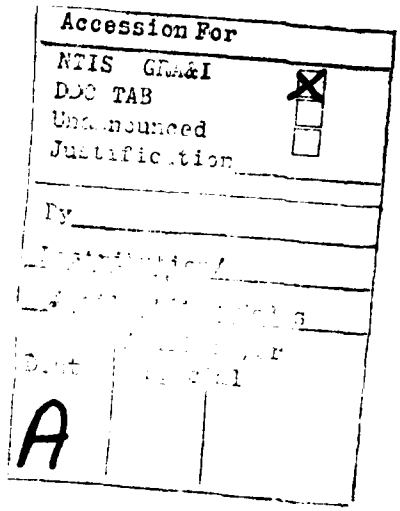
This paper is an introduction to some of the ideas of robust statistical methods as was presented to the Fourth International Congress for Mathematical Education, session on Exploratory Data Analysis. Most statistical methods taught and used today are very sensitive to bad or atypical data and can give meaningless results in their presence. Robust methods protect against these undesirable effects and can be incorporated into the teaching of statistics at all levels of complexity. We discuss the need for robust methods to supplement (not replace) standard procedures, suggest some considerations regarding teaching, and review some of the fundamental concepts of robust estimation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TEACHING ROBUST METHODS FOR EXPLORATORY DATA ANALYSIS



by

Andrew F. Siegel
Princeton University
and
Bell Laboratories

Technical Report No. 173, Series 2
Department of Statistics
Princeton University

October 1980

Supported in part by U.S. Army Research Office Contract DAAG29-79-C-0205, awarded to the Statistics Department, Princeton University.

TEACHING ROBUST METHODS FOR EXPLORATORY DATA ANALYSIS

by

Andrew F. Siegel
Princeton University

A B S T R A C T

This paper is an introduction to some of the ideas of robust statistical methods as was presented to the Fourth International Congress for Mathematical Education, session on Exploratory Data Analysis.

Most statistical methods taught and used today are very sensitive to bad or atypical data and can give meaningless results in their presence. Robust methods protect against these undesirable effects and can be incorporated into the teaching of statistics at all levels of complexity. We discuss the need for robust methods to supplement (not replace) standard procedures, suggest some considerations regarding teaching, and review some of the fundamental concepts of robust estimation.

Robust Methods for EDA

Robust methods, which protect against undesirable effects of unusual observations in the analysis of data, can easily be incorporated into the teaching of statistics at all levels. Because many of the basic concepts are simple, robustness can and should be discussed when the student is being introduced to statistical ideas.

Robustness should complement, not replace, standard statistical tools such as means, variances, least squares estimates, and other methods based on assumptions such as the normal distribution. In fact, many statisticians now recommend that a robust analysis be used routinely to help assess the validity of a more classical analysis, because hidden structure or problems with the data are often brought to light. If the classical and robust analyses approximately agree, this can be taken as a confirmation of the classical results by a secondary analysis. But when they disagree, there is work to be done because either errors in the data need to be corrected, or else unexpected structure remains to be discovered and explained.

The need for statistical robustness can be seen even in the basic problem of finding an "average" value to summarize a list of numbers. For example, to summarize the five numbers

$$7,\ 8,\ 6,\ 4,\ 100\ ,$$

the arithmetic mean is

$$\text{Mean} = \frac{7+8+6+4+100}{5} = 25$$

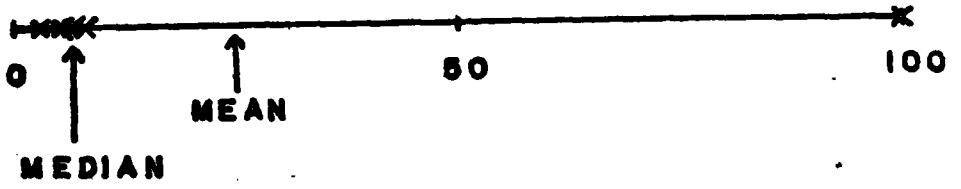
which is not a typical value! For some real-life problems, 25 would be the proper summary; but it is often better to

summarize the reasonable portion of the data (7,8,6, and 4) and to study exceptional values (like 100) separately, for example, to decide if they are interesting special cases for further study or simply in error.

The median is a robust measure of average which has half of the numbers smaller and half larger than itself. For this data set, it is

Median = Middle value of (4,6,7,8,100) = 7 ,

which we see is a typical value.



Robustness, formally, is protection against unusual data and violated assumptions. A few atypical or "bad" observations can ruin an ordinary analysis, but will have only a very limited effect on a robust analysis. Using robust methods is analogous to taking out an insurance policy for protection against the presence of bad data: the insurance premium is paid as an increase in sampling variation or efficiency of the estimate. In real data, errors are often present, and this "insurance" can be vital. Robust methods also help in the detection of outliers (atypical data), which can be very useful in error detection.

The teaching of robustness can proceed at many levels: simple or complex, pencil or computer, in-class or independent project. It can be taught separately as a section by itself, but is also easily integrated with other statistical topics.

For example, after teaching a new standard procedure, some time can be spent discussing methods of "robustifying" that method. The use of examples is crucial, of course, to teaching any statistical ideas and maintaining student interest; pictures and graphic displays should be used frequently.

To illustrate some robust methods for location (average) estimation, consider the attention spans of 10 hypothetical students:

5,18,15,2,8,55,11,3,9,8 minutes

The arithmetic mean (not robust) is

$$\text{mean} = \frac{5+18+\ldots+8}{10} = 13.4 \text{ minutes}$$

The 10% trimmed mean (robust) is formed by (1) ordering the data from smallest to largest, (2) trimming (removing) 10% of the data from each side, and (3) taking the arithmetic mean of what remains.

1) order: 2,3,5,8,8,9,11,15,18,55
2) trim 2 and 55
3) 10% timed near $= \frac{3+5+\ldots+18}{8} = 9.6$ minutes

The median (very robust against atypical values) is

$$\text{median} = \frac{8+9}{2} = 8.5 \text{ minutes.}$$

These estimators (mean, trimmed mean, median) are all examples of a rich family of location estimators called L-estimates, which are linear combinations of order statistics.

Another useful class that also includes robust members is

the M-estimates which generalize least squares and maximum likelihood procedures. These include the arithmetic mean, which minimizes the sum of squared deviations

$$\sum_{i=1}^{n} (x_i-\theta)^2 \quad .$$

In place of squaring, M-estimates allow a function ρ that can be less sensitive to outliers. We minimize

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

by differentiating and solving

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

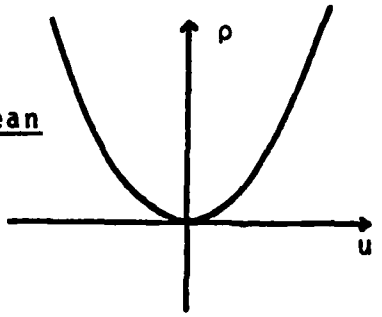
where ψ = (constant) $\cdot$ $(d\rho/d\theta)$. Different choices of ρ lead to different M-estimates with different properties. Some examples are given in Figure 1.

The median, an M-estimate with $\rho(x) = |x|$, is extremely resistant to bad data but suffers from "granularity", a lack of responsiveness to data near the central value. The Huber choice for ρ corrects this problem: near zero, it is like the mean, allowing data near the average to "fine-tune" the estimate, while maintaining resistance to bad data by behaving like the median away from the middle. Tukey's Bisquare also combines efficiency and robustness, but has a ψ that is "redescends" to zero; in effect, this says that data that are very far from the middle will not be believed, and will have zero effect on the estimate.

For easy pencil-and-paper calculation, L-estimates are preferable, because the minimization step for M-estimates (other

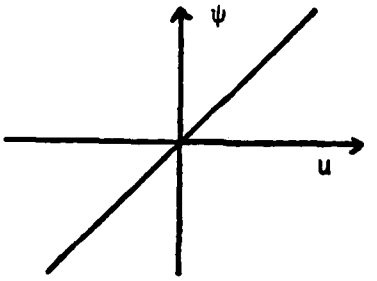
FIGURE 1. Examples of M-estimates

Mean

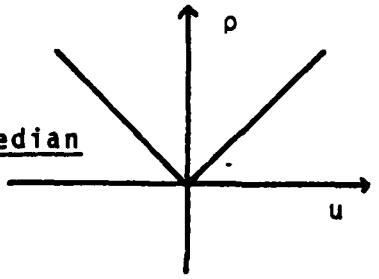


$$\rho(u) = u^2$$

$$\psi(u) = u$$

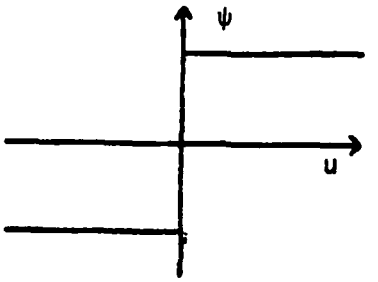


Median

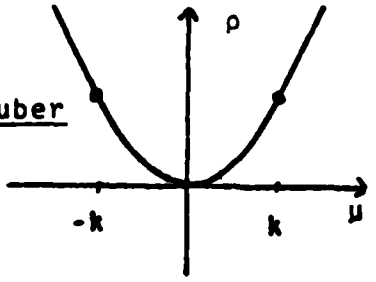


$$\rho(u) = |u|$$

$$\psi(u) = \begin{cases} -1 & u < 0 \\ 1 & u > 0 \end{cases}$$

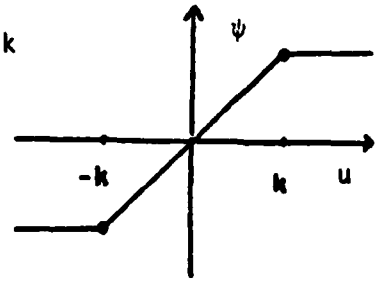


Huber

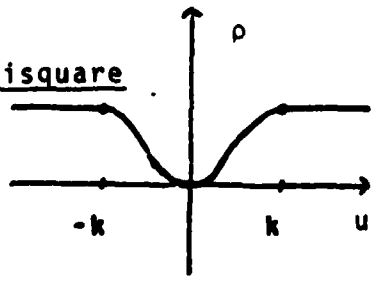


$$\rho(u) = \begin{cases} u^2/2 & |u| < k \\ ku - k^2/2 & |u| \geq k \end{cases}$$

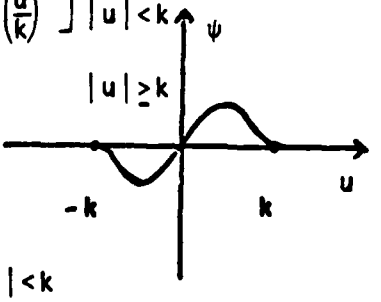
$$\psi(u) = \begin{cases} k & u > k \\ u & |u| \leq k \\ -k & u < -k \end{cases}$$



Bisquare



$$\rho(u) = \begin{cases} \frac{u^2}{6}\left[3 - 3\left(\frac{u}{k}\right)^2 + \left(\frac{u}{k}\right)^4\right] & |u| < k \\ k^2/6 & |u| \geq k \end{cases}$$



$$\psi(u) = \begin{cases} u\left[1 - \left(\frac{u}{k}\right)^2\right] & |u| < k \\ 0 & |u| \geq k \end{cases}$$

than the mean and median) is best attempted with a pocket calculator or computer.

The proportion of bad data that an estimation procedure can tolerate and still return a sensible answer is its Breakdown Value. The mean has a breakdown value of zero, because by changing the value of even a single number, the mean can be forced to assume any value as in Figure 2a. The median has a breakdown value of 50% because almost half of the data must be changed before the median breaks down completely, as illustrated in Figure 2b. Note that extreme observations do have an effect upon the median (compare the second and third parts of Figure 2b). Also note that when 3 of 5 points (more than 50%) are moved, the median breaks down as shown in Figure 2b. Breakdown Values of trimmed means lie in between those of the mean and median;for example, the 10% trimmed mean has a Breakdown Value of 10%.

FIGURE 2a. The mean has 0% breakdown value.

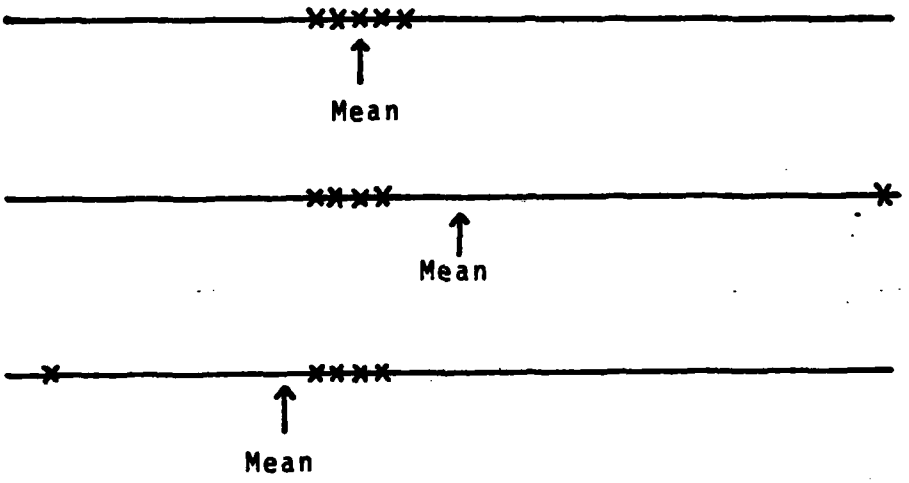
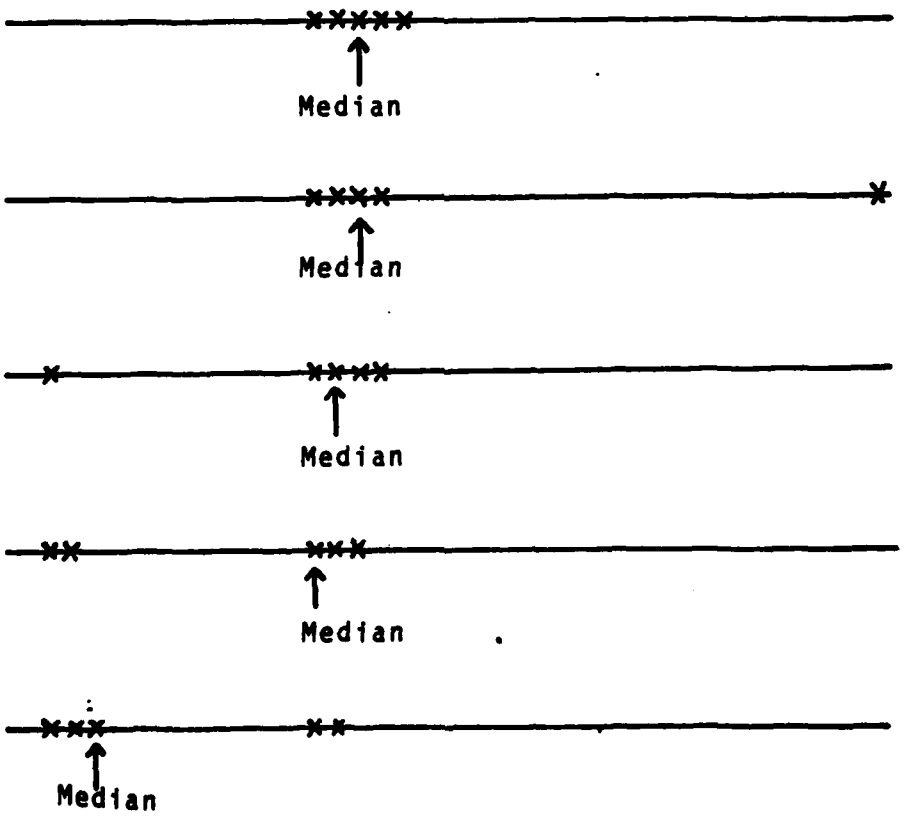


FIGURE 2b. The median has 50% breakdown value.



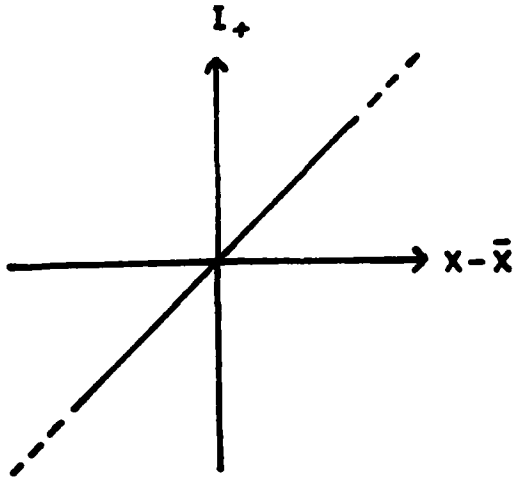
One measure of robustness of an estimate is provided by measuring the effect of adding a new point x to a sample $x_1,\ldots,x_n$. The Influence Function of the estimate $\hat{\theta}$ at the value x is defined to be

$$I_+(x,\hat{\theta}) = (n+1)\left\{\hat{\theta}(x_1,\ldots,x_n,x)-\hat{\theta}(x_1,\ldots,x_n)\right\}$$

For example, if $\hat{\theta}$ is the mean $(\Sigma x_i)/n$, we can calculate

$$I_+(x,\bar{x}) = x-\bar{x} .$$

Plotting I_+,



we see that the mean has an unbounded Influence Function, and is therefore not robust because there is no limit to the effect a single new point can have on the mean. For M-estimates, I_+ is very much like ψ.

Several alternatives exist for estimating scale that robustify the standard deviation:

$$S.D. = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

The "MAD" (Mean Absolute Deviation has the median) is obtained by replacing means by medians:

$$MAD = \text{Median}(|x_1 - m|, |x_2 - m|, -, |x_n - m|)$$

$$\text{where } m = \text{Median } (x_1, \ldots, x_n)$$

For example, an initial data set 7,8,6,4,100 has a SD = 42 but MAD = Median $(|7-7|, |8-7|, \ldots, |100-7|)$ = Median (0,1,1,3,93) = 1. The large standard deviation 42, is due to the fact that 100 is very far from most of the data set. The MAD, 1, is smaller because this single large contribution does not dominate.

Another robust scale estimate is the Interquartile Range, simply the upper quartile minus the lower quartile (after ordering the data, quartiles are 1/4 of the way in from each end).

Linear regression, fitting a straight line to points in two dimensions, can also be robustified, for example with M-estimation techniques. However, even M estimates can break down in a situation as in Figure 3. Which line do we want? The answer is

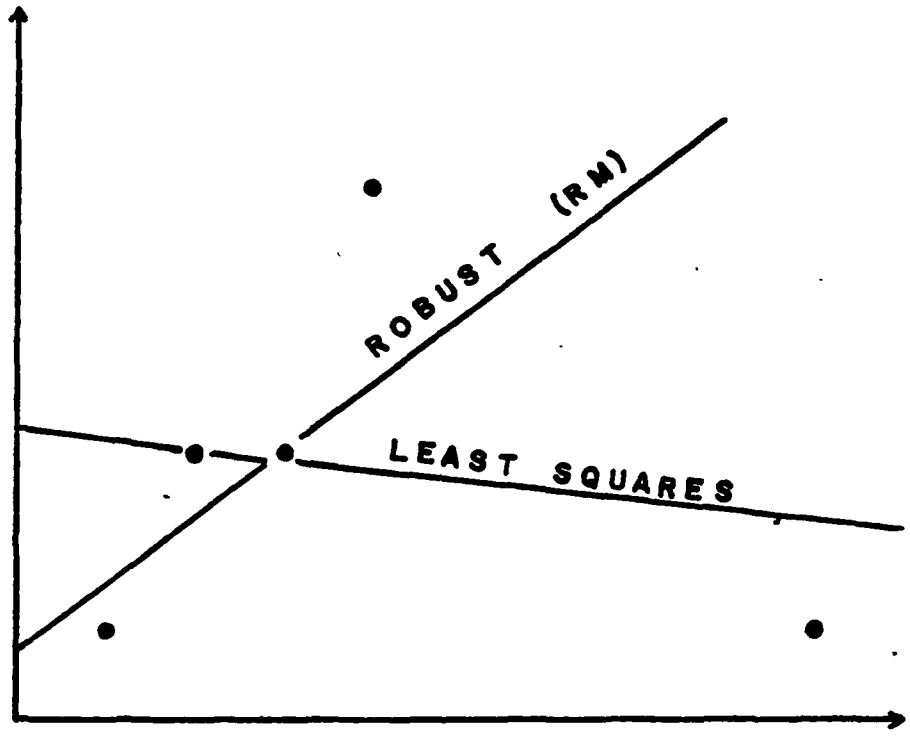
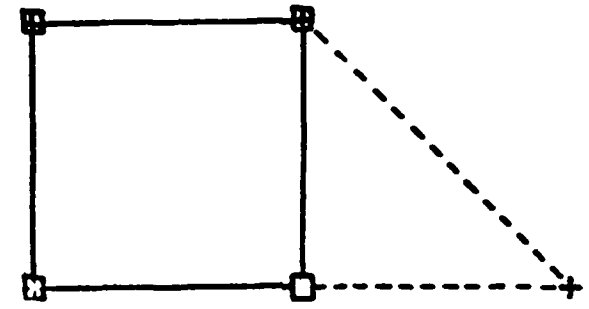


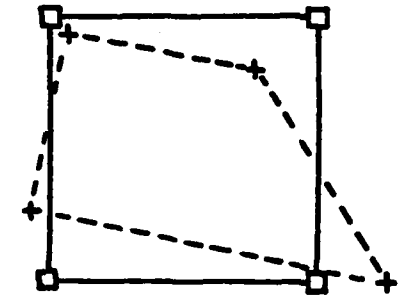
FIGURE 3. Leverage in Regression

"both lines". If the high-leverage point is in error, we prefer a robust line, such as the repeated median line (Launer and Siegel 1981). But the least squares line is preferable when that outlying point is correct, because in this case that single point provides nearly all our information about the slope!

A final example of the usefulness of robust methods is the fitting of two related shapes. Consider a square with one point distorted (dotted shape) fitted to a perfect square (solid shape) by allowing rotation, translation, and magnification. Figure 4 indicates the least squares fit, and the robust fit by Repeated Medians. Because the robust method "fits what fits" it indicates clearly that the dotted shape is identical to a square except at one point. The least squares fit, by compromising and trying to fit too much, makes this sort of inference much more difficult. Practical application of this type of shape-fitting has been demonstrated by Siegel and Benson (1980) in the comparison of fossil shapes and of human skulls.



Robust



Least Squares

FIGURE 4.

Robust methods are also available for correlation, time series, and two-way analysis in addition to the location, scale, and regression problems discussed here. For more information, we refer you to the reference list that follows. Remember that robustness is a young field (although its roots are deep in the past) and we can expect more books to become available in the near future.

ACKNOWLEDGEMENTS

I am grateful to R. Gnanadesikan, P. Tukey and J. Kettenring for helpful conversations during the preparation of this report.

REFERENCES

GENERAL

HOAGLIN, D.C., and WELSCH, R.E. (1978). The Hat Matrix in Regression and ANOVA. The American Statistician 32, 17-22.

HOGG, R.V. (1979). Statistical Robustness: One View of Its Use in Applications Today. The American Statistician 33, 108-115.

MALLOWS, C.L. (1979). Robust Methods - Some Examples of Their Use. The American Statistician 33, 179-184.

MC NEIL, D. (1976). Interactive Data Analysis. Wiley, New York.

TUKEY, J.W. (1977). Exploratory Data Analysis, Addison-Wesley, Reading, Massachusetts.

TUKEY, J.W., and MOSTELLER, F. (1977). Data Analysis and Regression, Addison-Wesley, Reading, Massachusetts. (See Chapter 10 for robustness.)

VELLEMAN, P.F., and HOAGLIN, D.C. (1980). Applications, Basics, and Computing of Exploratory Data Analysis. Duxbury Press, North Scituate, Massachusetts, forthcoming.

MORE TECHNICAL

ANDREWS, D.F. et al. (1972). Robust Estimates of Location. Princeton University Press.

HUBER, P.J. (1977). Robust Statistical Procedures. S.I.A.M. v. 56, Philadelphia.

LAUNER, R.L., and WILKINSON, G.N. ed (1979). Robustness in Statistics. Academic Press, New York.

LAUNER, R.L., and SIEGEL, A.F. ed (1981). Advances in Data Analysis. Academic Press, New York, forthcoming.

SIEGEL, A.F., and BENSON, R.H. (1980). Estimating Change in Animal Morphology. Submitted to Biometrics.