Discussion of:  Outliers by
R. J. Beckman and R. D. Cook

Discussants:  Charles E. McCulloch
Biometrics Unit
Cornell University
Ithaca, New York, 14853

and

Duane Meeter
Department of Statistics
Florida State University
Tallahassee, Florida, 32306
This paper represents a tremendous amount of work by the authors in trying to pull together results on what must be one of the most vexing and yet widespread of statistical problems. The coverage is broad with emphasis on outliers in normal linear models and an interesting section on the history of the subject. The latter material may carry an implied warning for present-day research.

Constraints of time and space have caused the authors (and us!) some problems, however. Some work is described so briefly that it is difficult to understand what was done, to say nothing of its implications (sec. 3.2, Hogg (1967); sec. 3.3, Kitagawa (1979); sec. 5.4, Andrews (1972)). Other work is evaluated so briefly that the reader may not be able to appreciate the reasoning involved. For example, the scheme of generating data with outliers by always modifying the extreme order statistics is labeled "an (unfortunate) attempt to generate observations what would usually be judged outlying according to the informal definitions of Edgeworth and Grubbs" in sec. 1.1. Later, in sec. 3.3, this approach is described as "making the studies useless" and it is also noted that in the normal case it "leaves both the underlying distribution and that of the outliers non-normal and does not mimic situations that occur in practice". We are not sure why the former is bad, and the latter criticism seems contradictory with the first quote, assuming Edgeworth and Grubbs were correctly describing a real phenomenon.

We contend that not enough seems to be known about what causes outliers, and this is one reason why there is little agreement about what to do with them. To quote George Box (1979), ". . . we cannot expect to get good results unless we are really prepared to engage in the hazardous undertaking of finding out what the world is really like." This is true of the world of outliers, too. In that spirit, we offer some comments from a hasty, informal survey of scientists,
asking questions such as: What causes outliers? What do you do with outliers? Do they usually tend in one direction?

**Psychologist, memory research:** Check protocol for failures in instructions or apparatus. If no assignable cause, do not delete the observation. Outliers almost invariably manifested as sub-par performance.

**Biochemist:** There must be a rational reason for deleting an outlier. High radiation counts suggest a dirty bottle; low counts, pipette too small or concentration too low, etc. In a series of preliminary experiments, an outlier may in retrospect be a valid indication of an unsuspected phenomenon.

**Meteorologist/Oceanographer:** Discard observations greater than 3σ automatically when estimating means in large data sets. Look for the physical explanation — recording, transmitting, computing errors. Maximum rainfall is determined by height of rain gauge; wind speed, by strength of wind gauge. A rainfall of zero is suspect in annual data; not in daily data.

**Biologist:** Check lab or field notes. Throw out data from sick animals. Most outliers are non-statistical. Data reflecting fitness have unidirectional outliers. Often multivariate temporal data on an animal reveal outliers on only one variable at or after a specific point, e.g., egg-laying.

**Physicist, high-energy research:** An outlier is a diagnostic symptom of a pathology in the experiment. Check equipment. All measurements have an allowable domain. For 99% of the time, something is wrong with the experiment. The other 1% of the outliers might represent progress. On automatic checking: Some variables must respond monotonically, within error; we check this automatically. In quark-hunting, perhaps one event in 100,000 is fiercely cared for, examined, and contended. Every experiment has its own peculiar dangers.
A word about terminology: Barnett and Lewis (1978, p. 22) use "outlier" where the authors use "discordant observation". But Barnett and Lewis (1978, p. 23) use "discordant observation" where our authors use "contaminant". May we suggest "data outlier" for the former, and "model outlier" for the latter? There is, of course, a parallel with the familiar concept of sample and population.

The authors suggest we need to standardize criteria for judging methods for outlier rejection. In the authors' terms, x and y are the numbers of true and false positive identifications of outliers, respectively. A bivariate distribution for x and y will in most applications be concentrated on the small non-negative integers. Do we really need the authors' six parameters to describe this distribution? (They don't justify their choices.)

We would have liked to have seen more on graphical approaches. Andrews (1972) got two sentences. Bradu and Hawkins (1982), one sentence, appears to be a promising technique, can be applied to higher-order tables than two-way, and contains definitions of masking and swamping that are easier to understand and more clearly located than those in the present work.

Much of the paper deals with methods appropriate for linear models. Linear model data sets typically contain many entries on variables interrelated in unknown ways. In such a circumstance, techniques involving personal scrutiny must inevitably miss many instances of dubious data. We commend the emphasis on recent research on outliers in normal linear models and regard it as the paper's main advantage over Barnett and Lewis (1978). However, we think it is optimistic to say that the mean shift model is equally valid for detecting an error in the ith row, $x_i^T$, of $X$ (as well as for detecting outlying responses $y_i$). In fact, one can only tell that the case is discrepant, not whether it comes from errors in independent or dependent variables. This leads to the thought that possible errors
in $X$ should likely be modeled differently when $X$ is a matrix of observational variables than when $X$ represents the levels in a designed experiment. In the latter case the error may be rare, but substantial. For example, setting a wrong experimental condition may involve only one or two entries; wrong instructions to one subject on one day may affect all of the entries in one row of $X$. Yet we use the $X$ which we thought was run, not the actual. In contrast, observational studies will produce a bewildering variety of errors. Most entries would be subject to measurement errors of various types. Consider for example the likelihood, sign and magnitude of measurement errors in "1981 income" and "sex of respondent".

The above reinforces the suggestion of Cook (1977) that the two components of

$$D_i = \frac{r_{ii}^2}{p} \frac{v_{ii}}{1-v_{ii}}$$

be analyzed separately as well as jointly. It is appealing that studentized residuals and leverage work together here, but is this combination appropriate as a general tool in outlier analysis? Perhaps one should look at $\log D_i$ and a scatterplot of $\log(r_{ii}^2)$ against $\log(v_{ii}/p(1-v_{ii}))$. Contours of constant $D_i$ are straight lines with slope $-1$; those who desire emphasis on residuals might use lines more horizontal, while increasing emphasis on "predictor outliers" suggests discrepancies should be measured relative to a more nearly vertical line.

We have made such a plot for the Longley data (see Figure 1) and have also plotted the 10% and 50% contours as suggested by Cook (1977) to give a rough guide to the magnitude of $\log D_i$. The plot shows that the points 1951 and 1962 correspond to the two highest leverage points, but not to the largest studentized residuals. This would be important to notice if one were interested in identifying outliers or assessing the influence of possible outliers on standard deviation.
estimates. If one were more interested in the influence of the possible outliers on the fitted regression, it would be appropriate to use the log $D_i$ contours. Finally we note that rough comparison contours could also be drawn in for log $r_i^2$ and log $\frac{1}{p} \frac{v_{ii}}{1-v_{ii}}$. For example, for leverage, using the rough guideline $\frac{2p}{n}$ for $v_{ii}$ (Hoaglin and Welsh, 1978) we could draw comparison contours at log $\frac{1}{p} \frac{v_{ii}}{1-v_{ii}} = \log \frac{2}{n-2p}$. For studentized residuals we could draw comparison contours at log $r_i^2 = \log \frac{(n-p)t^2}{t^2+n-p}$, where $t^2 = t_{\alpha/n}^2$ (Weisberg, 1980, Table D).

Finally, reading this survey and a number of the references has raised in us a concern that the study of outliers is becoming a research specialty. If we were to judge outlier research from other areas which have gone in this direction we might look for: narrowing of the scope of the investigation to the mathematically tractable cases (concentration in outlier research on the mean-shift and variance-inflation models?); ignoring the scientific origins of the problem ("... most recent papers lack even an informal definition"[of outlier]); excessive formalism (testing for a potentially damaging departure from assumptions using $\alpha = .05$); the appearance of surveys of the field (we jest); and a rise in the incidence of theorems and "optimal" results (so far, not too ominous).

We hope that this review article with its emphasis on historical foundations and definitions, as well as recent published results will encourage researchers in the field to get in touch with their roots. This seems essential in a field where the choice of technique depends so heavily on the assumed model.
REFERENCES


Figure 1 - Longley Function

\[ \log \frac{\nu_{1}}{p(1-v_{1})} \]

\[ \log F \]

Points marked with asterisks correspond to specific values of \( F \).
Figure 1 - Longley Data

\[ \log r^2_i \]

\[ \sim \log F_{9,50}^7 \]

\[ \sim \log F_{9,10}^7 \]

\[ \log \frac{\nu_{11}}{p(1-\nu_{11})} \]