This invited paper was written to mark the occasion of *Technometrics*'s 25th year of publication. The Editors and Management Committee would like to thank the authors and discussants for their efforts.

Outlier . . . . . s

R. J. Beckman
Los Alamos
National Laboratories
Los Alamos, NM 87545

R. D. Cook
Dept. of Applied Statistics
University of Minnesota
St. Paul, MN 55108

The concern over outliers is old and undoubtedly dates back to the first attempt to base conclusions on a set of statistical data. Comments by Bernoulli (1777) indicate that the practice of discarding discordant observations was commonplace 200 years ago. The history of the treatment of such observations is traced from these early comments to the present. For simple normal samples, M estimators, L estimators, and Bayesian estimators are presented as techniques that accommodate outliers in estimation. For the study of concomitant variables, methods are given for the outright rejection of discordant observations. Rejection techniques for multiple outliers are reviewed, as are the effects of masking and swamping. Methods for the detection of outliers in a regression setting are given for various models. Influence functions in multiple regression are discussed and compared with both the classical and the Bayesian methods of outlier detection. Work on outliers in circular data, discriminant analysis, experimental design, multivariate data, generalized linear models, distributions other than normal, and time series is also reviewed.

KEY WORDS: Bayesian outlier analysis; Design; Diagnostics; Discriminant analysis; Generalized linear models; Generalized residuals; Influential observations; L estimation; Masking; Multivariate outliers; Outlier history; Regression; Residual analysis; Robust estimation; Swamping.

CONTENTS

1. INTRODUCTION

The literature on outliers is vast and has much in common with many other areas—robust estimation, mixture models, slippage problems, data analysis in general, and so on—each of which is important in its own right. Indeed, the standard normal theory t test might be viewed as a test for a single outlier in a sample of size one. Here we give primary attention to the central developments and, where appropriate, briefly indicate the relationships to other problems.

The recent books by Barnett and Lewis (1978) and

Hawkins (1980) provide useful surveys of much of the relevant literature. Other works that contain substantial review material are Anscombe (1960), Barnett (1978), Grubbs (1969), Kale (1979), Prescott (1980), and Rider (1933). Stigler (1973,1975) gives an interesting historical account and Harter (1978) provides a pre-1950 annotated bibliography on order statistics that contains many references to outliers.

Here, we give a reasonably comprehensive account of the outlier literature with emphasis on methods that have been developed for application in the physical sciences.

In the second section of this article we survey the early history on outliers and show that, while the specific techniques for dealing with outliers have changed over time, the basic attitudes have not. In writing this section, we have made liberal use of quoted material since we found it difficult to paraphrase the early work while maintaining its essential flavor.

In Section 3 we survey the development of various methods for handling outliers in simple, normal random samples. This is basically an overview of the literature since details are readily available elsewhere; see, for example, David (1981), Hawkins (1980), and Barnett and Lewis (1978).

The major emphasis in this work is on outliers in linear models, as discussed in Section 4. This has been an active area for research in recent years, and many unsolved problems remain. We attempt to present a reasonably comprehensive and relatively detailed account of the literature, and to investigate the similarities between frequentist methods, Bayesian methods, and methods based on influence, as well as other selected approaches that have been proposed for dealing with outliers in linear models.

In Section 5 we discuss other problem areas, generalized linear models, designed experiments, multivariate samples, time series, and so on. The purpose of this section is to provide an indication of the state of the art in other areas by giving a brief introduction to the literature, suggesting possible solutions for selected unsolved problems, and indicating where further research may be required.

In the remainder of this section we discuss a variety of general ideas that we have found helpful when trying to sort through the overwhelming and often confusing literature on outliers. The ideas of an outlier and reasons for the study of outliers are discussed in Sections 1.1 and 1.2, respectively. In Sections 1.3 and 1.4 we discuss causes of outliers and general approaches to the problem.

1.1 What Is an Outlier?

No observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. An event with one chance in a million will occur with the appropriate frequency no matter how surprised we are that it should occur to us. Intuitively, the probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions. Observations that, in the opinion of the investigator, stand apart from the bulk of the data have been called "outliers," "discordant observations," "rogue values," "contaminants," "surprising values," "mavericks," and "dirty data," to mention only a few of the terms that have been coined over the years. Investigators are rightly concerned when such observations occur.

The concern over outliers is old and undoubtedly dates back to the first attempts to base conclusions on a set of statistical data. Comments by Bernoulli (1777) indicate that the practice of discarding discordant observations was commonplace 200 years ago. The first attempts to develop statistically objective methods for dealing with outliers were reported around 1850. It might be expected that a widely accepted, objective understanding of outliers and methods for dealing with them would have developed by now. This is not the case according to our interpretation of the literature. Although much has been written, the notion of an outlier seems as vague today as it was 200 years ago. For example, Edgeworth (1887) writes:

> Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined.

Eighty-two years later Grubbs (1969) states that

> An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs.

These statements illustrate that an outlier is a subjective, post-data concept. Historically, "objective" methods for dealing with outliers were employed only after the outliers were identified through a visual inspection of the data. To fix ideas, suppose that the following 10 observations (from Johnson and Hunt 1979) are intended to be independent realizations from a common normal population:

$$-1.64,\ -1.33,\ -1.10,\ -.57,\ -.27,$$
$$1.04,\quad 1.56,\quad 1.84,\quad 2.04,\quad 4.99 \qquad (1.1)$$

Which, if any, observations should we label as outliers? The response that we should reserve judgment until a test or some other objective measure has been constructed is a natural one, but the decision to employ an objective criterion is often based on initial, subjective reaction to the data. In this case, the observation 4.99 does seem too large, and some may question also the validity of the second largest observation, 2.04.

Collett and Lewis (1976) report the results of an

experiment to investigate the subjective nature of the decision to label an observation as an outlier. They conclude that an individual's willingness to perceive an outlier depends on the method of presentation (random, ordered, or graphical), on experience, and on the scale of the data; extreme observations tend to appear more discrepant as the scale is increased.

Further insight into the ways in which outliers are perceived can be gained by comparing the methods used to generate outliers in Monte Carlo studies of the performance of various procedures. In our review of the literature, we have noted two distinct schemes. In the first, a single outlier is generated by first obtaining a pseudorandom sample from the target population, usually standard normal, and then modifying a randomly selected member of the sample by adding, or multiplying by, a selected constant. In the second scheme the largest (or smallest) order statistic is always chosen for modification (Tiku 1975,1977; Jain and Pingel 1981b). The second scheme seems to represent an (unfortunate) attempt to generate observations that would usually be judged outlying according to the informal definitions of Edgeworth and Grubbs.

For the most part, the previous comments apply to small to moderate-sized simple random samples. In large data sets or more complicated settings—regression, multivariate samples, designed experiments, and so on—a worthwhile visual inspection of the data may be quite impossible. Thus it becomes necessary to apply some type of objective criterion at the outset, rather than after a visual inspection of the data. Many of these criteria are based on outlier models and thus some decisions about the nature and frequency of outliers are necessary before a criterion can be chosen. The subjectivity is introduced in the modeling stage and thereafter the resulting criterion is necessarily applied in a routine and objective manner.

Because of the emphasis on modeling in recent years, "outlier" now seems to be used by some authors to indicate any observation that does not come from the target population, although most recent papers lack even an informal definition. Of the 10 observations given at (1.1), 1.04, 2.04, and 4.99 were generated from a normal population with mean 3 and variance 1, $N(3,1)$, while the remaining 7 observations were generated from a $N(0,1)$ population. Some would say that this data set contains 3 outliers, although only one observation clearly stands out.

In an effort to avoid some ambiguity, we adopt the following working notions in the remainder of this article.

Discordant Observation—any observation that appears surprising or discrepant to the investigator.

Contaminant—any observation that is not a realization from the target distribution.

Outlier—a collective to refer to either a contaminant or a discordant observation.

1.2 Why Study Outliers?

One obvious answer to this question was indicated in the previous section: As long as investigators take the trouble to look at their data, discordant observations will occur. Clearly, an adequate method for handling such observations, or at least a firm understanding of the relative merits of available methods, is necessary. Although this is the dominant historical reason for the study of outliers, several additional and equally important reasons have been emphasized in recent years.

Special Interest. It may happen that interest centers on the discordant observation itself rather than on, for example, the estimation of a population parameter. In such situations the statistical problem involves drawing valid inferences about the observation in question.

Barnett (1978) describes the interesting legal case of Hadlum vs. Hadlum as an example of this type of special interest: 349 days after Mr. Hadlum departed for military service abroad, Mrs. Hadlum gave birth; Mr. Hadlum judged the observation of 349 days to be discordant when compared with the average gestation time of 280 days and therefore petitioned for divorce. In this case, the issue is not whether the discordant observation should be discarded or downweighted in estimation, but is how to judge the weight of evidence against the hypothesis that the discrepant observation is a valid (albeit extreme) realization from the distribution of gestation times.

Detection of Specific Alternative Phenomena. A few years ago the Russians inadvertently dropped one of their satellites in central Canada. The joint U.S. and Canadian effort to locate the debris involved monitoring changes in the level of radiation while flying over probable locations. Levels of radiation that were high relative to the background were taken as indicators of an alternative phenomenon, satellite debris. The basic statistical methodology consisted of applying outlier detection schemes on a large scale. There was no interest in estimating the average level of background radiation, except to provide a baseline for judging individual observations as outliers. (For a similar example, see Conover, Bement, and Iman 1979.)

Many examples of this type, where interest centers on detecting alternative, rare phenomena rather than on estimating characteristics of common, known phenomena, can be found in the scientific literature.

Diagnostic Indicators. In complicated analyses where a firm, well-grounded belief in the model is lacking, the presence of outliers is often an indication

of weaknesses in the model, the data, or both. Outlying observations in the original scale, for example, may conform when the responses are transformed to the logarithmic scale. (For further discussion of outliers and transformations, see Atkinson 1981,1982; and Carroll 1982). Highly influential observations in regression may be due to regions in the factor space with inadequate coverage, errors in the rows of the model matrix, or large residuals. Outlying observations may also reflect nonadditivity or heteroscedasticity.

Data analyses should include the application of selected diagnostic techniques that can furnish information on the appropriateness of the analysis and on the accuracy of conclusions. Many of the available techniques are based on detecting observations that are outlying in one respect or another (Cook and Weisberg 1982 provide a review).

Accommodation. The study of the nature and frequency of outliers in a particular problem can lead to models and estimation methods that accommodate outliers and thus that result in improved inferences. In such situations, outliers are usually of interest only insofar as they complicate the analysis. The parameters, if any, that describe the nature and frequency of the outliers are usually regarded as nuisance parameters.

Marks and Rao (1979) provide an example of what can be achieved in dental research through the study of outliers. Their study led them to conclude that a particular type of outlier that arises because of the subject's fatigue can be accommodated by a mixture of two normal populations with $\theta = .1$ as the known mixing parameter.

Robust methods can be regarded as omnibus methods of accommodation. These methods offer protection against certain general types of outliers, but do not recognize the specific relevant information that may be available in a particular problem.

Influence. FACE-2, the second phase of the Florida Area Cumulus Experiment, was designed to confirm the indication from phase one that cloud seeding can produce increases in natural rainfall. Kerr (1982) reports that the analysis of the FACE-2 data is complicated by a single outlier, an unseeded day on which the amount of rainfall was 4 standard deviations above the mean. Evidently, this single outlier is very influential: with the outlier the estimated increase in rainfall is about 5 percent (p value $\doteq .4$), while the estimated increase changes to 25 percent (p value $\doteq .13$) when the outlier is removed.

Clearly, outliers need not be influential; the results of an analysis may be essentially unchanged when an outlying observation is removed. It is useful to regard an influential observation as a special type of outlier. The study of such observations enables one to obtain a better understanding of the structure of a problem. For example, such studies may uncover inherent deficiencies in the data and lead to further experimentation.

1.3 Causes of Outliers

Several specific causes of outliers are suggested in the previous section and good discussions can be found in the review papers listed previously. For our purposes it is sufficient to arrange the causes into three broad and somewhat overlapping categories: global model weaknesses, local model weaknesses, and natural variability. As implied by the first two categories, outliers must be judged with some model, either implicit or explicit, in mind. From a statistical view, it is useful to think of the appearance of an outlier as being due to the inability of the model to provide an adequate fit or statistical explanation. Global model weaknesses are those causes that would lead to the replacement of the present model with a new or revised model for the entire sample. This category includes causes such as response variables in the wrong scale or frequently occurring outliers of a known nature. The former cause may lead to a transformation of the response, while the latter cause may lead to replacing the present model with a mixture, as in Marks and Rao (1979).

Local model weaknesses are those causes that apply to only the outlying observations and not to the model as a whole. Such causes may require that the outliers be dealt with individually, since an alternative model that incorporates the outliers will usually be unknown. Examples include isolated measurement and recording errors, and highly influential observations in regression due to remote points in the factor space.

Finally, an outlier may be the result of natural variation rather than any weakness of the model. Neyman and Scott (1971) introduced the concept of outlier-prone and outlier-resistant families $\mathcal{F}$ of distributions in an effort to characterize the extent to which discordant observations will be naturally produced. Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the order statistics of a sample from a distribution $F \in \mathcal{F}$. Then $\mathcal{F}$ is said to be "(K, n) outlier resistant" if

$$P(K, n) = \sup_{F \in \mathcal{F}} \Pr[y_{(n)} - y_{(n-1)} > K(y_{(n-1)} - y_{(1)}) \mid F] < 1.$$

Otherwise, $\mathcal{F}$ is said to be "(K, n) outlier prone." If $P(K, n) = 1$ for every $K > 0$ and $n \geq 3$, $\mathcal{F}$ is said to be "completely outlier prone." Presumably, our behavior in the presence of discordant observations should depend on the type of family, but we know of no statistical application for these concepts. In fact, ac-

cording to the working definitions given in Section 1.1, a "completely outlier prone" family produces outliers (discordant observations) only in the absence of information about the family. Once we know that the family is "outlier prone" remote observations should no longer appear discordant.

For further work along these lines, see Green (1974,1976).

1.4 Identification and Accommodation

Following Barnett (1978) and Barnett and Lewis (1978), we distinguish between two broad methods for dealing with the possibility of outliers. The first method is simply to identify the outliers for further study. Identification of an outlier may lead to (a) its subsequent rejection, (b) important new information contained in concomitant variables that would otherwise have gone unnoticed, (c) its incorporation through a revision of the model (global weaknesses), or method of estimation, or (d) a recognition of an inherent weakness in the data and thus to further experimentation. Diagnostics based on the identification of outliers are particularly useful during the iterative construction of a model where a critical assessment is necessary at various stages of development; general discussions are given by Box (1979,1980) and Cook and Weisberg (1982).

The second method is to accommodate the possibility of outliers by suitable modifications of the model and/or method of analysis. For example, mixture models can be used to accommodate certain types of contaminants and M estimates are often used to provide some protection against outliers when the mixture model is symmetric.

Of these two general methods, we judge identification to be the more fundamental. Methods of accommodating outliers tend to require much information about the process generating the outliers or are designed to be immune to the presence of outliers. The latter types often obscure essential information, while the former types require that much essential information be already obtained, usually through the process of identification. Although the philosophies underlying the notions of identification and accommodation seem distinct, they are often confused, since a method of accommodation may produce a method of identification as a by-product, or vice versa.

In small- to moderate-size univariate samples, a visual inspection of the data is perhaps the most common method of identifying outliers. Many ad hoc tests are available to lend "objective" statistical support. The concept of an outlier, while vague, is best understood in this situation: observations on the extremes of the sample that are well separated from the bulk of the data are usually given special attention.

In more complicated settings, the notion of an outlier becomes more elusive. A general idea behind methods to identify a single outlier is to transform the data into a set of n univariate statistics, one for each observation in the data, that are inspected visually or, in some cases, used to construct significance tests. Many different transforms have been suggested; some are based on specific alternative models and some are designed to reflect specific characteristics of the individual observations.

For example, the Studentized residuals are often used to identify outliers in analyses based on linear models. The observation with the largest absolute Studentized residual is usually given special attention and is taken as the observation most likely to be a contaminant. For justification of this procedure many authors appeal to the assumption that the basic normal theory model is valid except that the expectation of at most one unknown response is shifted; in effect, a single parameter is devoted to a single observation. The resulting *mean-shift* model has dominated the literature on outliers in linear models.

In contrast, the *variance-inflation* model for a single outlier is based on the assumption that the basic model is valid except that the variance of one unknown response is larger than the rest. The analysis of the variance-inflation model tends to be more complicated, but Cook, Holschuh, and Weisberg (1982) show that the observation most likely to have an inflated variance need not correspond to the observation with the largest absolute ordinary residual $\max |e_i|$ or the largest absolute Studentized residual $\max |r_i|$. However, if the observation that corresponds to $\max |e_i|$ also corresponds to $\max |r_i|$, then the mean-shift and variance-inflation models agree with respect to the most outlying observation.

One point worth special emphasis is that methods of identification and accommodation that are based on alternative models can produce results that are specific to the model in question; thus the determination of an outlier can depend critically on our understanding of the problem. In particular, it is useful to distinguish between situations where the outliers are assumed to carry some information about the parameters in the basic model and situations where the outliers carry absolutely no relevant information. The variance-inflation model is an example of the former sort of situation, while the mean-shift model is an example of the latter.

Thus far we have considered only methods of identifying outliers that are dependent on the specification of an alternative model. In recent years a number of outlier diagnostics have been proposed that do not require such an alternative. When robust regression is viewed as iteratively reweighted least squares, the weights from the final iteration may be useful indicators of outliers (Hogg 1979). Similar comments apply

to bounded-influence regression (Krasker and Welsch 1982). Cook (1977,1979) proposes a criterion for detecting the observations that have a substantial influence on the least squares estimate of the parameter vector in linear regression. Cook's proposal is one of a number of recent criteria that have been suggested for measuring the influence of individual or groups of observations on selected phases of an analysis. Andrews and Pregibon (1978) propose a criterion for "finding the outliers that matter." The criteria proposed by Cook (1977) and Andrews and Pregibon are compared in Draper and John (1981). Johnson and Geisser (1980a,b) study methods for assessing the influence of observations on predictive distributions. Pregibon (1981) considers influential observations in logistic regression, and Atkinson (1981,1982) proposes methods for detecting observations that influence the selection of a transformation for the response variable. A comprehensive treatment of much of the literature on influence is available in the recent books by Cook and Weisberg (1982), and Belsley, Kuh, and Welsch (1980).

This recent work is a step away from the traditional approach to outliers since it is based on a specialization of purpose. An influential observation is clearly one that is outlying in terms of its influence on the particular phase of the analysis that is being monitored.

2. EARLY HISTORY

The earliest discussion of outliers that we have been able to locate is by Bernoulli (1777). Bernoulli evidently questioned the assumption of identically distributed errors and condemned the widespread practice of discarding discordant observations in the absence of a priori information:

> But is it right to hold that the several observations are of the same weight or moment, or equally prone to any and every error? Are errors of some degrees as easy to make as others of as many minutes? Is there everywhere the same probability? Such an assertion would be quite absurd, which is undoubtedly the reason why astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability. This practice makes it more than clear that they are far from assigning the same validity to each of the observations they have made, for they reject some in their entirety, while in the case of others they do not only retain them all but, moreover, treat them alike. I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others. Nevertheless, I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations. If there is no such reason for dissatisfaction I think each and every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken every care.

Judging from the comments of Peirce (1852), the practice of rejecting discordant observations was still commonplace some 75 years after Bernoulli, but objective criteria for justifying the practice still had not been developed:

> In most every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve, in the present state of science, to perplex and mislead the inquirer. Geometers have, therefore, been in the habit of rejecting those observations which appeared to them liable to unusual defects, although no exact criterion has been proposed to test and authorize such a procedure, and this delicate subject has been left to the arbitrary discretion of individual computers.

There is a stormy history behind the rejection of outliers. In the past, as is the case today, the lines were fairly well drawn between those who discarded discordant observations, those who gave each observation a different weight, and those who used simple, unweighted averages.

The first person to put forth a specific criterion for the rejection of outliers was Peirce (1852,1878). He proposed to solve the problem of discordant observations using the idea, which is called "bizarre" by Barnett and Lewis (1978), that

> The principle upon which it is proposed to solve this problem is, that the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more, abnormal observations.

Three years later Gould (1855) supplied tables to help in the implementation of Peirce's criterion. In doing so he stated that

> Professor Peirce has given the results of the successful investigation of a singular problem, and one unquestionably among the most important of any which could be proposed in its relations to all those exact sciences to which quantitative research or measurement may be applied. ... The delicate task of discriminating between such observations, and those whose discordance, although great, ought not to be deemed abnormal, has hitherto been left to the arbitrary judgement of individuals ...

However, the praise that Gould gave to the technique was not universal. The noted professor of astronomy Airy (1856), possibly motivated by an error in the published scale of longitudes for England caused by the retention of only the "most accordant observations," totally rejected Peirce's idea:

> And I have, not without surprise to myself, been led to think, that the whole theory is defective in its foundation, and illusory in its results; that no rule for the exclusion of observations can be obtained by any process founded purely upon a consideration of the discordance of those observations.

Airy argued from the premise that, while some observations were unlikely, any observation was possible and no observation should be preferred to another if there were "no cogent reason for supposing that unusual causes of error must have intervened in special observations." Like Bernoulli, he advised that the

mean of the observations be taken, for "we are bound to admit all on the same terms as giving equally valid evidence." Realizing that the philosophy of never rejecting any observation no matter how large the error is contrary to common sense, he then rejected the mathematical theory of normal errors rather than the observation.

Wintlock (1856) found Peirce's criterion both correct and of value. He chastised Airy for tending "in no unimportant degree, to retard the general adoption of the 'Criterion,' especially among that class of astronomers who would not be sufficiently attracted to it ... to take the trouble to satisfy themselves of the soundness of the theoretical basis on which it rests." Wintlock argued that Airy accepts, a posteriori, the magnitude of discordant observations when using the method of least squares. Wintlock then wondered

> Why should there be, *a priori*, any greater objection to approaching a series of observations with a theory of antecedent probability for the purpose of detecting abnormality in some of them, than for the purpose of ascertaining any other peculiarity or property that they may possess, whose existence or magnitude has to be received or rejected by us on an estimate of the probabilities for or against it?

Chauvenet (1960, written in 1863) gave both his own and Peirce's criterion for the rejection of a single outlier. He derived a critical point χ so that the expected number of residuals larger than χ is less than $\frac{1}{2}$. Chauvenet then concluded that an error of that magnitude or greater "will have a greater probability against it than for it and may therefore be rejected."

A third proposal of the outright rejection of discordant observations was made by Stone (1868). He found the use of Peirce's criterion to be troublesome and, citing Airy, expressed doubts about the correctness of the mathematics. He failed to note, however, that Airy's main objection to the Peirce criterion was to the entire philosophy of discarding discordant observations. Stone also objected to the Chauvenet method since he found it to be based upon an "erroneous principle." He reasoned that Chauvenet, by discarding (on the average) one observation in every two data sets, does not take into account "the average number of observations that (a) person makes with one mistake," which he called the "modulus of carelessness." Stone's procedure then was to reject all observations whose probability of occurrence is less than the modulus of carelessness.

While the techniques of outright rejection of discordant observations of Peirce, Chauvenet, and Stone generated much controversy, they did not seem to be widely used. Saunders (1903) called attention to Peirce's criterion: "especially as I believed that the criterion is seldom if ever employed." Also, these authors worked with the model of Gaussian errors rather than with mixtures of Gaussians, which would have made their arguments for the outright rejection of the outliers more convincing. Under the single Gaussian model, the authors called discordant those observations that must by necessity have come from the parent distribution.

One of the first persons to postulate differing distributions for the observations was Glaisher (1873,1874). He assumed that each observation came from a Gaussian distribution with unequal variances, and using a scheme of iterative reweighted least squares, obtained an estimate for the mean.

Glaisher's work was a conceptual breakthrough in the area of outlier treatment for two reasons. First, it is very much like robust estimation in that less weight was given to outlying observations. In addition, Glaisher was the first person to hint that the observations may have come from different populations. However, Glaisher is probably remembered most for his scathing comments about the methods of Stone and Peirce rather than for his reweighting scheme. In Glaisher (1873) fully one-half of the manuscript contains comments about Stone's method. He did not find Peirce's criterion unsound in a mathematical sense, but "because I thought the principle of rejecting an observation in toto (except in the case of an obvious mistake) unsound." About Stone, however, he wrote: "Mr. Stone's criterion, however, does not appear to be good even among criteria [for rejecting observations in toto]." He goes on to say

> In the first place it is to be remarked that the criterion cannot be intended as a practical one, as it is quite out of the question to make even a rough guess at the value of [the modulus of carelessness]; for (1) we do not know what sources of error are included, nor (2) within what limits the error is to be detected and corrected with perfect or almost perfect certainty; and even were these two "unknowables" given, the determination of [the modulus of carelessness] would be sufficiently arbitrary.

Stone (1873) replied to Glaisher with some scathing comments of his own. He defended his own technique and found the Glaisher method not only not "mathematically complete" as Glaisher claimed, but also "mathematically unsound."

Many of the differences between these two authors may have arisen owing to different underlying models, for in this latter paper Stone hinted at a mixture model whose components have differing means: "There are such things as systematic errors recognized in practice, i.e. errors which, for an observation made with a particular instrument in a particular way, will always enter into the result with the same sign."

The Stone-Glaisher disagreement ended abruptly in 1874 when Glaisher was forced to withdraw a paper from the Monthly Notices of the Royal Astronomical Society. Instead he published a 20-line note (Glaisher 1874) in which he stated: "considering the tone of Mr. Stone's remarks ... I felt myself justified in expressing with frankness what I thought." He goes on to state that, "It was suggested to me that the paper contain-

ing these opinions—necessarily somewhat personal—should be withdrawn, and I was advised that the matter rest where it was, without further comment, a course which is, perhaps, the better."

Iterative weighting of observations was put on a more sound basis by Newcomb (1886). Stating that "it is nearly always found that some of the outstanding errors seem abnormally large," Newcomb found two objections to the Peirce criterion for dealing with these abnormal observations. His first objection was that the procedure took no account of the possible known errors, but rather determined the error in an observation by its relationship with the other observations. His second objection was that no account was taken of the fact that "the probability that an observer should make an abnormal observation varies with the observer." (Newcomb does not mention Stone's work.) Newcomb goes on to state that any outright rejection scheme leads to a "discontinuous function of the separate errors of observation." He postulated a "mixed system of observations" of changing variances, in which the mean value is estimated by weighting the observations. One of the remarkable things is that the weights that he obtained correspond to robust estimation with a ψ function of $\psi(x) = \max(-c, \min(x, c))$ which is claimed by Stigler (1973) to be one of "Huber's favorite M-estimators."

Edgeworth (1887) hypothesized three models for outliers and illustrated each using Monte Carlo methods. These models were then used as a basis for a comparative study of the methods of Airy, Chauvenet, Glaisher, Newcombe, Peirce, and Stone. He (a) found Stone's modulus of carelessness to be legitimate, (b) referred to Airy's suggestion as the "No-Method," (c) recognized the different models underlying the work of Stone and Glaisher, (d) indicated that the criteria of Chauvenet and Peirce may be too pessimistic, and (e) expressed high regard for Newcomb's approach: "Professor Newcomb soars high above the others, in that he alone ascends to the philosophical, the *utilitarian*, principles on which depends the whole art of reducing observations."

3. OUTLIERS IN SIMPLE NORMAL SAMPLES

3.1 Introduction

In the early 20th century, the multiple lines of investigations of methods for dealing with discordant observations continued. Some authors still advocated the outright rejection of observations, some the use of weights, while others advocated the development of new models to incorporate such observations. Irwin (1925) and Student (1927) favored outright rejection, with Student reasoning that "Many if not most routine analyses may have a leptokurtic error system ... and in such cases rejection of outlying observations improves the accuracy of the mean." Daniell (1920), on the other hand, encouraged the use of weights based on the sampling distribution, with weights decreasing with the size of the deviation from the mean for "super normal" distributions (heavy-tailed), and increasing to the boundaries for "subnormal" (light-tailed) distributions. The notion of a new model for the incorporation of discordant observations was put forth by Fisher (1922). Much like Airy of the previous century, Fisher stated, "as a statistical measure, however, the rejection of observations is too crude to be defended: and unless there are other reasons for rejection than mere divergence from the majority; it would be more philosophical to accept these extreme values, not as gross errors, but as indications that the distribution of errors is not normal."

Authors in the 19th century, with the notable exception of Glaisher, were accustomed to thinking of simple normal sampling distributions. In the 20th century, however, many different models appeared to explain the presence of discordant observations. As mentioned previously, Student, Daniell, and Fisher all postulated the use of heavy-tailed distributions. Later in the 20th century, Ogrodnikoff (1928), Jeffreys (1932), Dixon (1950), and Grubbs (1950) suggested mixture models to describe sampling distributions that produce outliers. Dixon (1950) proposed two types of mixtures: The first of these was a mean-shift model (called model A by Ferguson 1961) where the distribution is a mixture of $N(\mu, \sigma^2)$ and $N(\mu + \lambda, \sigma^2)$ components, while the "variance-inflation model" (called Model B by Ferguson 1961) is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, a^2\sigma^2)$ components, $a^2 > 1$.

The use of such models led to the consideration of formal statistical tests for the rejection of outlying observations. Interestingly, while most authors cite improved estimation and accommodation as reasons for rejecting outliers, the development of formal test statistics is more in line with the rejection of outliers for the purposes of special interest and the detection of specific alternative phenomena. Still, the development of tests for the rejection of outliers continues today. Most authors are rather vague about the use of such tests and ignore the warnings of Collett and Lewis (1976) about the interpretations of the α level.

The development of tests for the rejection of outliers led to a widening of the gap between those who advocate rejecting outliers and those who prefer a less severe weighting scheme. Most proponents of tests now do not mention estimation, while those interested in weighted means or robust estimators pay little attention to the rejection of outliers for the study of concomitant variables.

The purpose of this section is to trace the history of outlier detection and accommodation techniques in simple normal samples. As the gap in the literature

between weighting schemes and outright rejection widens, we leave the history of robust estimation and trace only the development of the rejection techniques, since methods for dealing with outliers are presently synonymous with such techniques. This is not to say that rejection techniques should be preferred universally. On the contrary, we think that robust estimation is one of the best ways to accommodate outliers in estimation problems, and we encourage its routine use. For the most part, detailed discussions of techniques that have been extended to linear models are reserved for Section 4.

The remainder of this section is organized into two parts: The first deals with methods for the accommodation of outliers in location problems, and the second deals with techniques that are designed for the outright rejection of outliers.

3.2 Accommodation

The idea of weighting observations to obtain a robust estimator of the mean carried over into the 20th century, and there developed four main lines of attack. The first of these, L estimation, involves weighting the order statistics. Stigler (1973) gives an account of the early history of such estimators. In a second method, the weights of the observations are based on the residuals. M estimators fall into this class. Bayesian methods constitute a third class. Finally, we discuss a hybrid method consisting of outright rejection followed by estimation (weights = 0 or 1) with the emphasis on the performance of the estimator.

Methods Based on Order Statistics. As the name indicates, L estimators are linear estimators in the order statistics. If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the order statistics then an L estimator of a location parameter is given by $T = \sum_{i=1}^{n} a_i X_{(i)}$, where the a_i's are appropriate weights. These types of estimators include the median and various trimmed and Winsorized means where the values of the a_i may depend on the sample.

Daniell (1920) measured the relative efficiencies of various L estimators in terms of ratios of mean squared errors. Among the estimators Daniell investigated were optimal weight functions g, with weights equal to $(1/n)g(i/(n+1))$, and the trimmed mean that he called the discard average.

L estimators were used by Sarhan and Greenberg (1956) for naturally censored samples, but Dixon (1960) recognized that censoring may also be imposed as a method for dealing with outliers. He showed the efficiency of Winsorized estimators, as measured by the ratio of their variance to that of the best linear systematic statistic of Sarhan and Greenberg, to be at least .94.

The virtues of Winsorized means were expounded by Tukey (1960) on the grounds that long-tailed distributions are more common than short. Using a variance-inflation model, he showed that with long-tailed distributions the median is more efficient than the sample average as an estimator of the mean with as little as 10 percent contamination.

Hodges and Lehmann (1963) used optimal rank tests to obtain an estimator (not an L estimator) for the mean using the median of all pair averages of the data; that is $T = \text{median}\,(\frac{1}{2}(X_i + X_j), 1 \le i \le j \le n)$.

Bickel (1965) compared the efficiencies of the Hodges-Lehmann estimator, the trimmed mean, and the Winsorized mean. He found the Hodges-Lehmann estimator to be "preferred in any situation where the degree of contamination and type of distribution is not known with great precision."

A simple estimator given by $.3F_n^{-1}(\frac{1}{3}) + .4F_n^{-1}(\frac{1}{2}) + .3F_n^{-1}(\frac{2}{3})$, where F_n is the empirical distribution function, was proposed by Gastwirth (1966). Asymptotic theory of various L estimators has been investigated by Bickel and Hodges (1967), Dixon and Tukey (1968), and Jaeckel (1971a). For further discussion of L estimators, see Serfling (1980, Ch. 8).

Estimators Based on Residuals and Moments. The Glaisher-Newcomb idea of iteratively reweighting observations based on the size of their residuals seems to have disappeared from the literature until 1964. This is probably due to the inordinate amount of time needed to determine the basic estimates for even the smallest of sample sizes and to the ease with which estimators such as the trimmed mean could be computed. In a remarkable paper, Huber (1964) established a new approach to the theory of robust estimation. This approach is based on the notion that the poor performance of the sample average in the presence of contamination is due to the sensitivity of the least squares objective function $\sum (X_i - T)^2$ to outliers. This sensitivity can be reduced and bounded by replacing the least squares objective function with $\sum \rho(X_i - T)$, where ρ is a suitably selected loss function. This class of M estimators contains the sample mean with $\rho(t) = t^2$, the sample median with $\rho(t) = |t|$, and the maximum likelihood estimator with $\rho(t) = -\log f(t)$. By measuring robustness as the supremum of the asymptotic variance of the estimator as the underlying distribution ranged over all mixtures $(1 - \varepsilon)\Phi + \varepsilon H$ for a fixed ε and symmetric H, Huber showed that the most robust estimator corresponds to choosing

$$\begin{aligned} \rho(t) &= t^2/2 && |t| < K \\ &= K|t| - K^2/2 && |t| \ge K. \end{aligned}$$

Huber (1983) discusses three different approaches to M estimation, including minimax interval estimates

that provide protection against small amounts of asymmetric contamination.

Hogg (1967) invented an adaptive robust estimator based on the kurtosis of the sample and the assumption that the true distribution is one of a specified family of m distributions $\{D_1, D_2, \ldots D_m\}$, each with mean θ and kurtosis k_j, with $k_i \neq k_j$ for $i \neq j$. Assume further that there exists a preferred estimator T_i of θ for each distribution D_i, and that there are m disjoint intervals $I_1, I_2, \ldots, I_m$ so that k_j is an interior point of $I_j, j = 1, 2, \ldots, m$. Then the estimator is given by T_j if k, the sample kurtosis, is contained in the interval I_j.

Since Huber's pioneering work in 1964 there has been an enormous amount of literature on robust estimation. For further information on the subject the reader is directed to Andrews et al. (1972), Hogg (1974), Huber (1972,1977,1981), Hampel (1974), and Jaeckel (1971b).

Bayesian "Weights." Using Newcomb's idea that each observation may come from a Gaussian distribution with a different variance, Ogrodnikoff (1928) placed a prior distribution on the variance and weighted the observations by the ratio of the posterior third moment to the posterior mean.

Jeffreys (1932) considered a mean-shift model where the variances of the two populations and the mixing parameter are known. He assumed uniform prior distributions for the two mean values and estimated the posterior means by "successive approximations."

The posterior mean was proposed by de Finetti (1961) to reduce the influence of outlying observations. De Finetti stated that "according to the Bayesian point of view, there exist no observations to be rejected" and that "the final distribution (of the unknown mean) is determined on the ground of all the observations taken."

The use of Bayesian estimators to minimize the weighted sum of risks for a quadratic loss function was proposed by Gebhardt (1964). Assuming that there is at most one contaminant in the data, he considered both the mean-shift and variance-inflation models and viewed the problem as a multiple-decision one, with the prior probability that the ith observation is an outlier being given.

Finally, Dempster and Rosner (1971) assumed that there are k contaminants in the sample with location shifts of unknown magnitude. By placing a uniform prior over the $\binom{n}{k}$ possible configurations of k outlying observations, they were able to find the posterior probability that a particular configuration was the one containing the k outliers. They propose that this be done for $k = 0, 1, \ldots$, and the model that best fits the data be chosen.

For further discussion of outliers and Bayes inference in simple, normal samples, see Goldstein (1982), de Alba and Van Ryzin (1980a,b), O'Hagan (1979), Guttman and Kahatri (1975), Guttman (1973a), and Box and Tiao (1968).

Rejection Techniques Used in Estimation. Techniques for the outright rejection of outliers are discussed in Section 3.2. However, since some authors have considered how these rejection rules influence the estimation of the mean, we consider these studies here. Most of the investigations of the behavior of the estimated mean rely on the mean squared error. For example, Dixon (1953) compared the mean squared error of the mean and the median for various types of contamination. In addition, he investigated the mean squared error of the mean after outlying observations were removed by a gap test that we describe in Section 3.2. He found that the mean squared error for the gap procedure is smaller than that of the median for small sample sizes.

In contrast to Dixon's work, Anscombe (1960) developed procedures for the rejection of outliers that are directly dependent upon the mean square of the estimator. Citing the work of Chauvenet and Peirce, Anscombe questioned the use of significance tests in outlier studies: "Most modern statisticians, bemused by 5%, give rules (for the rejection of outliers) having rejection rates of about one per $20n$ observations. No one has explained why this should be so. No one seems to have asked. Rejection rules are not significance tests." Anscombe then introduced the idea that rejection rules should be treated like insurance policies and hence should be judged according to the premium one has to pay and the protection given by the use of such a rule. The premium of a rule is the fractional increase in mean squared error that results when there are no contaminants present and the protection is the corresponding decrease in mean squared error when contaminants are present. Anscombe proposed three rules for the rejection of outliers and subsequent estimation of the mean. Each of these techniques involves rejecting observations according to the size of their standardized residuals and the premium one is willing to pay. The premium and protection for these rules were investigated for samples of size three.

Tiao and Guttman (1967), Guttman and Smith (1969,1971), and Guttman (1973b) continued to promote premium-protection rules for estimation of the mean. When the variance is known and there is at most one spurious observation in the sample, Tiao and Guttman propose the estimator

$$\hat{\mu} = \bar{X} - w,$$

where

$$w = \begin{cases} 0 & \text{if } |z_i| < c \text{ for all } i \\ (X_i - \bar{X})/(n-1) & \text{if } |z_i| > c,\ |z_j| < |z_i| \\ & \text{for all } j \neq i. \end{cases}$$

The variables z_i, however, are not the residuals, but are, curiously, the residuals with the same random normal added to induce independence. Values of c for various sample sizes, premiums, and protections were given.

Using both the mean-shift and the variance-inflation models, Guttman and Smith (1969) studied three rejection rules. These were Anscombe's rule, a Winsorized mean, and what they termed a semiwinsorized mean, where the outlying observations are replaced by a bound rather than the next largest or smallest observation. Similar rules were given by Guttman and Smith (1971) for estimation of the variance in the presence of possible outliers. In both of these papers, constants for fixing the premium were given by Monte Carlo for values of n up to 11. The protections, however, were investigated only for samples of size three.

Guttman (1973b) extended his previous work and gave complex rules to guard against more than one outlying observation. Constants for the use of his rules were obtained by Monte Carlo for sample sizes up to 100, and an extensive study of the protection of each rule was given. Further studies of the premium and protection were given by Marks and Rao (1978), where the number of contaminants was assumed to follow a binomial distribution.

3.3 Identification

Many formal techniques for the identification of outlying observations have been proposed. While some authors viewed these techniques as a way to improve the estimation of the mean, we prefer to think of them as tests that help us in our understanding of concomitant variables. This subsection is divided into two parts. The first deals with the detection of one or two outlying observations. Masking, swamping, and the testing of multiple outliers are covered in the second part.

Identification of One or Two Outlying Observations. The early work of Peirce, Chauvenet, and Stone on formal tests for the rejection of outlying observations was continued in the 20th century. Irwin (1925) proposed the use of the gap between the first and second, and the second and third order statistics as a test statistic for the rejection of outliers. By rejection he meant "the realization of the fact that the particular observations in question probably do not belong to the same homogeneous group." He gave critical values for these statistics based on a known variance. In a similar vein, Tippett (1925) proposed the use of the range for the detection of outliers. Once again he assumed that the variance is known. McKay (1935) took a different approach and found the distribution of the difference between the extreme observation and the sample mean scaled by a known standard deviation, that is, $(X_{(n)} - \bar{X})/\sigma$. The distribution of these variables was later tabled by Nair (1948), who also tabled the distribution of $(X_{(n)} - \bar{X})/\hat{\sigma}_v$, where $\hat{\sigma}_v$ is an independent estimator of σ.

Noting that in practice σ is estimated by the sample standard deviation, Thompson (1935) derived the distribution of an arbitrary Studentized residual, $\tau_i = (X_i - \bar{X})/\hat{\sigma}$. He showed that $(n-2)^{1/2}\tau_i/(n-1-\tau_i^2)^{1/2}$ has a Student's t distribution with $n-1$ degrees of freedom. Fixing the frequency of rejection per sample, Thompson obtained critical values.

Pearson and Chandra Sekar (1936) pointed out that while Thompson's criterion is both useful and practical for the detection of one outlier, the Studentized residuals are not independent and that in practice the experimenter would look only at the Studentized residual with the largest absolute value. By using the distribution given by Thompson, they were able to obtain critical values for $\tau = \max \tau_i$ for a limited number of sample sizes and without deriving its exact distribution.

Grubbs (1950) derived an expression for the exact distribution of τ. For testing for one outlier he proposed the use of the statistic S_n^2/S^2, where $S^2 = \sum^n (X_i - \bar{X})^2$, $S_n^2 = \sum^{n-1} (X_{(i)} - \bar{X}_n)^2$, and $\bar{X}_n = \sum^{n-1} X_{(i)}/(n-1)$. Grubbs showed that $S_n^2/S^2 = 1 - \tau^2/(n-1)$, hence S_n^2/S^2 is statistically equivalent to the maximum absolute Studentized residual. For testing that the two largest or the two smallest observations are outliers, he proposed and tabled the percentage points for the statistics $S_{n-1,n}^2/S^2$ and $S_{1,2}^2/S^2$, where $S_{n-1,n}^2 = \sum^{n-2} (X_{(i)} - \bar{X}_{n,n-1})^2$, $\bar{X}_{n,n-1} = \sum^{n-2} X_{(i)}/(n-2)$, and $S_{1,2}^2$ is defined in a similar manner. Grubbs, however, was unable to obtain critical values for the test statistic $S_{1,n}^2/S^2$ for testing simultaneously the largest and smallest observations as outliers. The procedure for testing for exactly one outlying observation was shown to maximize the probability of making the correct decision under the mean-shift model by Murphy (1951), Paulson (1952), and Kudo (1956). Ferguson (1961) showed the same when the outlier is generated according to the variance-inflation model. However, as we shall see later, this is not the case when multiple outliers are considered.

King (1953) extended Grubbs test to two-sided alternatives, while Quesenberry and David (1961) used a pooled estimator of the standard deviation for the denominator of the Studentized residual.

Murphy (1951) gave a location/scale-invariant procedure for maximizing the probability of a correct decision when there is contamination by mean-shift. The test statistic for the largest k out of n observations as outliers is of the form

$$T = \left(\sum_{i=0}^{k-1} X_{(n-i)} - k\bar{X} \right) \Big/ S.$$

This, of course, is Grubb's test when $k = 1$. Tests of this type, where sets of data are simultaneously tested as discordant are called *block tests* by Barnett and Lewis (1978).

While Grubbs was extending the work of Thompson, and Pearson and Chandra Sekar on Studentized residuals, Dixon (1950) proposed the use of the ratios of ranges and subranges to test for one or two outliers. In rather confusing notation, he gave statistics such as

$$r_{10} = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}$$

for the single outlier X_1, and

$$r_{20} = \frac{X_{(3)} - X_{(1)}}{X_{(n)} - X_{(1)}}$$

for outlier X_1 avoiding X_2. The distributions of ratios of the form $(X_{(n)} - X_{(n-j)})/(X_{(n)} - X_{(i)})$ for small values of i and j are given in Dixon (1951), and the performance of various procedures was investigated in Dixon (1962).

For observations assumed to come from a symmetric distribution, Walsh (1950) proposed the use of a nonparametric test to decide whether the r largest observations came from a population with the same median as the remaining observations. This test can be viewed as a gap test similar to those of Dixon.

Ferguson (1961) considered various invariant tests for outlier rejection. He showed that tests based on the sample skewness are locally best invariant tests for outliers with small positive shifts in the mean and that tests based on the sample kurtosis are locally best invariant tests for testing for small shifts in the mean in either direction. Tests based on the sample kurtosis are also locally best invariant tests for testing small positive shifts in the variance. Although tests based on either kurtosis or skewness are locally optimal, no test has much power in determining outliers for small shifts in the mean. In addition, Ferguson showed that for samples of size $n = 25$ the Studentized residuals are more powerful for detecting intermediate to large (greater than 3σ) shifts in the mean.

The repeated applications of single outlier tests for the rejection of multiple outliers is called a *consecutive test* by Barnett and Lewis (1978). McMillan and David (1971) and McMillan (1971) conducted power studies comparing consecutive versions of maximum Studentized or standardized residuals with Murphy's block procedure for the detection of two outliers. When the variance is known, McMillian and David showed that Murphy's test does only slightly better than repeated applications of the normal residual test. However, McMillan showed that when the variance is unknown Murphy's procedure slightly outperforms Grubbs's test for two outliers, but the consecutive Studentized residual test was shown to decrease in power as the two outlying observations moved further from the mean in the same direction. This decrease in power for consecutive tests is called *masking*.

Masking, Swamping, and Multiple Outlier Tests. That the power of consecutive applications of Grubbs's test for outliers decreases as the two outliers move further from the mean did not surprise McMillan (1971); Pearson and Chandra Sekar (1936) had shown that, for small samples, two or three outlying observations placed approximately the same distance from the mean so inflate both the sample mean and the variance that it is impossible to obtain a large value of the test statistic, $(X_{(n)} - \bar{X})/\hat{\sigma}$. Pearson and Chandra Sekar state "...it would appear that the criterion is only really efficient in the presence of a single outlying observation."

By using plots of the rejection regions of the consecutive Grubbs procedure, McMillan (1971) showed that if the outliers have exactly the same value they are never rejected. Further, if the outliers have exactly the same expectation, the probability that they are rejected goes to zero as their expectation increases. Prescott (1978) refined the rejection-region plots of McMillan by proposing plots of the sensitivity of the test statistics as a function of the outlying observations X_1 and X_2. The sensitivity function was obtained by evaluation of the test statistic at the $n - 2$ expected values of the normal order statistics from a sample of size $n - 2$ and the values of X_1, X_2. Contours of the sensitivity function were then plotted in the X_1, X_2 plane. For samples of sizes $n = 20$ sensitivity contours for consecutive use of the single outlier procedures of Grubbs, the r statistics of Dixon, and the skewness of Ferguson all have large regions where one outlier masks another. The sample kurtosis, however, is relatively free of masking. Hawkins (1980) pointed out that the masking effect of two outliers tends to decrease as the sample size increases, but as the sample size increases the number of outliers probably increases and hence more masking occurs.

Numerous writers, realizing that consecutive application of single outlier tests leads to problems, proposed tests for the simultaneous rejection of outliers. The first proposed test was the optimal block test of Murphy (1951). Tietjen and Moore (1972,1979) proposed two Grubbs-type statistics for the detection of multiple outliers. For testing k outliers on one side of the data they proposed

$$L_k = \sum_{i=1}^{n-k} (X_{(i)} - \bar{X}_k)^2/S^2,$$

where $\bar{X}_k = \sum_{i=1}^{n-k} X_{(i)}/(n-k)$. For k outliers on either side they gave

$$E_k = \sum_{i=k+1}^{n-k-1} (z_{(i)} - \bar{z}_k)^2/S^2,$$

where $z_{(i)}$ is the value of X corresponding to the ith largest (or smallest) ordered absolute residual. Although k was assumed known in the formal development, Tietjen and Moore suggest that k be estimated by use of the largest gap in the order statistics when testing for outliers on one side of the data. They give no indication of how to estimate k for tests based on E_k. Hawkins (1979) showed that E_k, even for the correct choice of k, can pick the wrong observations as outliers if the outliers occur on one side of the data. The flaw in the E_k statistic was corrected by Hawkins (1979). Reasoning that a severe outlier contaminates the mean, he used a procedure first proposed by Rosner (1975). In this method the mean is recomputed after the most severe outlier is removed and then the next most severe outlier is judged by its deviation from the new mean. This procedure is continued until all the data are ranked.

A test procedure for the simultaneous rejection of k outliers was also proposed by Tiku (1975). He gave as a test statistic the ratio of the maximum likelihood estimator of σ from a censored sample to that from an uncensored sample. The amount of censoring is estimated by the maximum gap in the order statistics, scaled by its expected value. (For more on estimating the amount of censoring see Jain and Pingel 1981a). The performance of this test in practice has not yet been suitably investigated. Tiku did a power study, but he contaminated only the largest (or smallest) observations generated from a normal distribution. This leaves both the underlying distribution and that of the outliers nonnormal and does not mimic situations that occur in practice.

In all of the previous tests for blocks of outlying observations, the experimenter is faced with the estimation or selection of k, the number of outliers. In a significant paper, Rosner (1975) developed test statistics that are not prone to masking and require only knowledge of the maximum number of possible outliers, k_u. Using Rosner's notation, the tests are constructed as follows: Consider the subsets $I_0, I_1, \ldots, I_n$, where $I_0 = \{x_1, x_2, \ldots, x_n\}$ and I_{t+1} is formed by deleting from I_t the point x^t farthest from the mean of the points in I_t. Next, let $R_t = S(I_{t-1})$, where $S(I)$ is any statistic for detecting one outlier. Rosner determined constants $\lambda_i(\beta)$ such that

$$\Pr\left\{\bigcup_{i=1}^{k_u}[R_i > \lambda_i(\beta)]\right\} = \alpha.$$

The test procedure declares $x^0, x^1, \ldots x^{l-1}$ as outliers, where $l = \max\{i: R_i > \lambda_i(\beta)\}$. If $R_i < \lambda_i(\beta)$ for all i, then no outliers are declared. Rosner studied the power of four outlier statistics: the extreme Studentized deviate, the Studentized range, the kurtosis, and an R statistic that is the maximum deviation from a trimmed mean scaled by a trimmed standard deviation. The test procedure based on the extreme Studentized deviate seems to be the most powerful.

Extended critical values for Rosner's technique are given by Rosner (1977), Prescott (1979), Chhikara and Feiveson (1980), and Jain (1981a). Power studies of the Rosner procedure were carried out by Jain (1981b) and Jain and Pingel (1981b), but in each of these studies the Tiku method of contamination was used, making the studies useless.

Hawkins (1978) and Cook and Beckman (1980) show that, while the Rosner procedure controls the α level for the rejection of the null hypothesis, it tends to declare more outliers than there are in the sample when the null hypothesis is rejected. This phenomenon is known as *swamping* (Fieller 1976) and is characteristic of many procedures for detecting multiple outliers when k_u is chosen larger than the true number of contaminants present.

Historically, the most annoying issue was how to choose k or an estimate thereof. If k was selected a priori and was smaller than the number of contaminants present, we ran the risk of not detecting any contaminants because of masking. On the other hand, if k was too large we may have declared valid observations as outliers because of swamping. Rosner's recommendation is to specify a clear upper bound k_u for k and then to employ some type of sequential procedure based on identifying the most outlying subsets on some criterion for $k = 1, 2, \ldots, k_u$. The choice of k_u is difficult and usually left to the discretion of the investigator, but k_u/n is typically visualized as being quite small. Nevertheless, the problems in choosing k_u are similar to those encountered by choosing k. In particular, if k_u is much larger than the true number of outliers, masking may not pose much of a problem but swamping will.

In an attempt to circumvent the problems of swamping and having to specify k_u, Cook and Beckman (1980) proposed the use of M estimators for the detection of outliers. Viewing Huber's M estimators as iteratively reweighted least squares, they obtained critical values for the weights. In a power study, they showed that this procedure outperforms the Rosner procedure in those cases where k_u is not "close" to the true number of outliers.

Finally, for the detection of multiple outliers Kitagawa (1979) used Akaike's information criterion,

$$\text{AIC} = -2 \log (\text{maximum likelihood}) + 2 (\text{number of independently adjusted parameters}),$$

which is an approximation of minus two times the expected entropy. By defining a series of models where the smallest r_1 observations belong to one outlying population and the largest r_2 to another, the model

with the minimum AIC (maximum entropy) was determined, and the corresponding r_1 smallest and r_2 largest observations declared outliers. Kitagawa gives numerous examples comparing this procedure with those of Grubbs, Dixon, and Tietjen and Moore.

3.4 Remarks

The number of articles on outliers has increased dramatically in the last few years owing to recognition of the problems presented by multiple outliers. If the number of outliers is known to be at most one, then almost all outlier techniques choose the observation with the largest Studentized residual as the most likely outlying observation. However, when multiple outliers are present, most techniques suffer from either masking or swamping or both. This is possibly due to the use of the magnitude of the residuals (or some variant of the residuals) as the test statistic. For multiple outliers from a univariate sample, the best test procedures may be those that fit distributions to the entire sample, such as Kitagawa's (1979) use of entropy or the many Bayesian techniques. An intensive power study of these types of procedures could be very useful. One of the main tasks in such studies will be the development of fast algorithms, since the methods are often computationally infeasible.

Techniques developed for the accommodation of outliers, such as M estimators and trimmed means, seem to do well with data generated according to the variance-inflation model. More thought should be given to techniques that accommodate outliers from the mean-shift model.

As a last remark, methods for judging the performance of various techniques need to be standardized. It is clear that the mean squared error is useful for evaluating accommodation methods. Methods for rejection, however, present a larger problem. Masking, swamping, and the number of observations correctly identified need to be studied for these methods. Letting x be the number of outliers correctly identified and y be the number of nonoutliers that are declared outliers (the number of false positives), we suggest the following six parameters be estimated:

1. The probability that no false positives and all outliers are detected, $\Pr(x = k, y = 0)$;
2. The probability of no false positives and at least one outlier, $\Pr(x > 0, y = 0)$;
3. The expected proportion of detected outliers, Ex/n;
4. The probability of at least one false positive, $\Pr(y > 0)$;
5. The expected number of false positives, $E(y)$; and
6. The proportion of false positives among all observations that were declared outliers, $Ey/(Ex + Ey)$.

4. OUTLIERS IN NORMAL LINEAR MODELS

4.1 Introduction

In this section we discuss various ways that have been proposed for extending the treatment of outliers in simple normal samples to the linear model

$$\mathbf{Y} = (y_i) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\mathbf{Y}$ is an n-vector of observable responses, $\mathbf{X}$ is a known, full-rank, $n \times p$ matrix, $\boldsymbol{\beta}$ is a p-vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an unobservable n-vector of independent and identically distributed (iid) normal errors with mean $\mathbf{0}$ and variance σ^2. We first describe what might be viewed as the central frequentist approaches to the detection of outliers in analyses based on (4.1). We next consider related but alternative approaches, including Bayesian treatments of the problem, and finally we briefly consider the relationship between influential observations and outliers.

4.2 Central Frequentist Approaches

Let $\mathbf{u}_i$ denote an n-vector with a 1 in the ith position and zeros elsewhere. Much of the work on outliers in linear models is based, sometimes implicitly, on the mean-shift model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{4.2}$$

where $\boldsymbol{\gamma}$ is a k-vector of unknown parameters, $\mathbf{D}$ is an $n \times k$ matrix with columns $\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_k}$, and the remaining quantities are as defined in (4.1). Srikantan (1961) and Ferguson (1961) were the first to explicitly state this model as a basis for dealing with outliers in linear models, although this approach is implicit in Daniel's (1960) work on two-way tables and was used extensively in prior investigations of outliers in simple random samples. This model is most appropriate when identification rather than accommodation is of interest.

The mean-shift model is usually stated as a basis for detecting outlying responses y_i, but it is equally valid for detecting an error in the ith row $\mathbf{x}_i^T$ of $\mathbf{X}$. Suppose, for example, that $k = 1$ and the observed $\mathbf{x}_i = \mathbf{x}_i - \boldsymbol{\Delta}_i$, where $\boldsymbol{\Delta}_i$ is unknown. Then the resulting model will be of the form of (4.2) with $\gamma = \boldsymbol{\Delta}_i^T\boldsymbol{\beta}$. Following Cook and Weisberg (1982), we refer to the vector $(y_i, \mathbf{x}_i^T)$ as the ith *case*. The mean-shift model is best regarded as a basis for detecting outlying cases, particularly in regression problems where the elements of $\mathbf{X}$ may not be verifiable.

Barnett and Lewis (1978) distinguish between the *labeled* version of (4.2) in which $\mathbf{i} = \{i_1, \ldots, i_k\}$ is assumed known prior to an inspection of the data and the *unlabeled* version in which this set of indices is unknown. Since the labeled version is a standard linear model, the analysis can be carried out using the

usual methods; and the results can be usefully expressed in terms of the analysis of the null model (4.1).

Labeled Model. Let $\hat{\boldsymbol{\beta}}$ denote the ordinary least squares estimator of $\boldsymbol{\beta}$ under (4.1) and let

$$\mathbf{V} = (v_{ij}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{4.3}$$

denote the $n \times n$ matrix that projects $\mathbf{Y}$ onto the vector of fitted values, $\hat{\mathbf{Y}} = (\hat{y}_i) = \mathbf{VY}$. The n vector of ordinary residuals under (4.1) is

$$\mathbf{e} = (e_i) = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{V})\mathbf{Y} \tag{4.4}$$

and the usual estimate of σ^2 is $\hat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/(n-p)$. Clearly, $\mathbf{e}$ follows a singular multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\mathbf{I} - \mathbf{V})$, so that $\text{var}(e_i) = \sigma^2(1 - v_{ii})$. Finally, the subvector of $\mathbf{e}$ and the submatrix of $\mathbf{V}$ that correspond to the cases indexed by $\mathbf{i}$ will be denoted by $\mathbf{e}_\mathbf{i}$ and $\mathbf{V}_\mathbf{i}$, respectively.

The least squares estimate of $\boldsymbol{\gamma}$ in the labeled version of (4.2) is simply

$$\hat{\boldsymbol{\gamma}} = [\mathbf{I} - \mathbf{V}_\mathbf{i}]^{-1}\mathbf{e}_\mathbf{i}, \tag{4.5}$$

provided, of course, that the indicated inverse exists. The least-squares estimates of $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}}_{(\mathbf{i})}$ and $\hat{\sigma}^2_{(\mathbf{i})}$, the usual estimates of β and σ^2 computed from (4.1) after discarding the cases indexed by $\mathbf{i}$. Since the analysis of the mean-shift model in terms of $\boldsymbol{\beta}$ and σ^2 is equivalent to discarding the suspect cases, it seems fair to regard this model as one in which the contaminants furnish no information about the basic process. John and Draper (1978), John (1978), and Draper and John (1980) carry the notion of discarding cases further. They note that the mean-shift model is similar to that used when estimating missing values with an analysis of covariance; thus they regard the *responses* indexed by $\mathbf{i}$ as missing values to be replaced by "estimates." Although the answers are the same, approaching the problem in terms of missing values seems more a hindrance than a help.

The additional reduction in the sum of squares due to fitting $\boldsymbol{\gamma}$ after $\boldsymbol{\beta}$ is

$$Q_k(\mathbf{i}) = \mathbf{e}_\mathbf{i}^T[\mathbf{I} - \mathbf{V}_\mathbf{i}]^{-1}\mathbf{e}_\mathbf{i}. \tag{4.6}$$

This notation is a slight modification of that introduced by Gentleman and Wilk (1975b), who provide many of the basic details for an analysis of the labeled version. Under the null model (4.1), $Q_k(\mathbf{i})$ can be expressed as a sum of squares of iid normal variables by using the spectral decomposition $\mathbf{V}_\mathbf{i} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is a $k \times k$ orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues $0 \le \lambda_1 \le \cdots \le \lambda_k < 1$:

$$Q_k(\mathbf{i}) = (\boldsymbol{\Gamma}^T\mathbf{e}_\mathbf{i})^T(\mathbf{I} - \boldsymbol{\Lambda})^{-1}(\boldsymbol{\Gamma}^T\mathbf{e}_\mathbf{i}) \tag{4.7}$$

The elements of $(\boldsymbol{\Gamma}^T\mathbf{e}_\mathbf{i})(\mathbf{I} - \boldsymbol{\Lambda})^{-1/2}$ are iid $N(0, \sigma^2)$ random variables and are constructed in a manner similar to the "successively adjusted residuals" of John and Draper (1978); see also Gentleman (1980) and Cook and Weisberg (1982). When $k = 1$, $Q_1(i) = e_i^2/(1 - v_{ii})$ is proportional to the square of the ith Studentized residual

$$r_i = e_i/(1 - v_{ii})^{1/2}\hat{\sigma}, \qquad i = 1, 2, \ldots, n. \tag{4.8}$$

The Studentized residuals play a central role in the detection of a single contaminant. Under the null model, the marginal distribution of $r_i^2/(n-p)$ is Beta with parameters $\frac{1}{2}$ and $(n - p - 1)/2$, $E(r_i) = 0$, $\text{var}(r_i) = 1$ and $\text{cov}(r_i, r_j) = -v_{ij}/[(1 - v_{ii})(1 - v_{jj})]^{1/2}$. The joint distribution of the r_i has been investigated by Ellenberg (1973) and Ferguson (1961); see also Doornbos and Prins (1958) and Joshi (1972).

The usual F statistic for testing the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ is

$$F_k(\mathbf{i}) = (n - p - k)Q_k(\mathbf{i})/k(e^Te - Q_k(\mathbf{i})) \tag{4.9}$$

with k and $n - p - k$ degrees of freedom. When $k = 1$, this reduces to (see Beckman and Trussell 1974)

$$F_1(i) = r_i^2(n - p - 1)/(n - p - r_i^2) \tag{4.10}$$

which is monotonically increasing in r_i^2. Srikantan (1961) established that when $k = 1$ the uniformly most powerful unbiased test of the hypothesis $\gamma = 0$ is based on the Studentized residual corresponding to the suspect case. Generally, it is worthwhile remembering that the power of this test can vary greatly depending on the structure of $\mathbf{V}$. The noncentrality parameter associated with the test of the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ is $\boldsymbol{\gamma}^T(\mathbf{I} - \mathbf{V}_\mathbf{i})\boldsymbol{\gamma}/2\sigma^2$, which reduces to $\gamma^2(1 - v_{ii})/2\sigma^2$ when $k = 1$. The term $(1 - v_{ii})$ will be small for remote points $\mathbf{x}_i$ in the factor space and thus outliers at such points will be the most difficult to detect (Cook 1979; Huber 1975; Davies and Hutton 1975). Generally, $\mathbf{V}$ and in particular its diagonal elements play an important role in the study of outliers. For a review of many relevant properties of $\mathbf{V}$, see Hoaglin and Welsch (1978), Cook and Weisberg (1980,1982), and Belsley, Kuh, and Welsch (1980).

The analysis of the labeled version is relatively straightforward. If necessary, standard subset-selection methods can be used to sort out probable zero elements of $\boldsymbol{\gamma}$; see Schweder (1976) for further suggestions. The analysis of the unlabeled version, however, is much more complicated since $\mathbf{i}$ and usually k are unknown. For a fixed k, a common recommendation is to search over all $\binom{n}{k}$ subsets and estimate $\mathbf{i}$ using the set of indices $\hat{\mathbf{i}}$ that maximizes $Q_k(\mathbf{i})$, although many approximate versions of this idea have been suggested in an effort to avoid costly calculations. This is equivalent to identifying the subset of k cases that result in the greatest increase in the maximized likelihood. In what follows we first

assume that $k = 1$ and later turn attention to situations in which $k > 1$.

Unlabeled model, $k = 1$. The use of the maximum absolute Studentized residual $|r_m| = \max |r_i|$, or a statistically equivalent criterion, for dealing with a single outlier in simple normal samples can be traced back to the work of Pearson and Chandra Sekar (1936). Anscombe (1960) and, in the same issue of *Technometrics*, Daniel (1960) were the first to propose the use of the maximum Studentized residual in connection with linear models. Anscombe's (1960) approach is based on using the now widely recognized concepts of premium and protection to characterize the performance of rules for the rejection of observations prior to estimation, as discussed in Section 3. Anscombe's application of these ideas to linear models is restricted to (a) situations in which the residuals under the null model (4.1) have constant variance so that correct weighting is not an issue and (b) rules of the form: reject the observation corresponding to $|e_m| = \max |e_i|$ if $|e_m| > C\sigma$ where C is a selected positive constant. Anscombe justified the use of residual-based rules on the grounds that they are natural and convenient. In this approach, Anscombe considered the effects of rejecting cases on the subsequent estimates—a concern that is conspicuously absent in many of the approaches based on the mean-shift model—rather than the validity of the cases in question. This line of development is extended by Guttman and Dutter (1976).

Daniel (1960) considered the problems of detecting and testing for a single "biased value" in an otherwise additive two-way table. He argued that, in the presence of a single outlier, the correlation between the residuals and their biases will be high and that this correlation should be used to detect the most probable biased cell. This line of investigation led Daniel to conclude that "... all the information about a single bad value is in the largest residual, $d_{\max}$, after all." Since the residual variances are constant in a balanced two-way table, there is no essential distinction to be made between the ordinary and the Studentized residuals.

Srikantan (1961) recommended the use of r_m^2, the square of the largest Studentized residual, as a test statistic for a single outlier in the unlabeled linear model, and was evidently the first to suggest that approximate critical values be based on the first Bonferroni bound; see Joshi (1972) for analogous results when σ^2 is known or an independent estimate is available. Ferguson (1961) formulated the problem in a decision-theoretic context with the available actions being represented by the $n + 1$ hypotheses H_0: $\boldsymbol{\gamma} = \mathbf{0}$, H_i: $\boldsymbol{\gamma} = \gamma_i \mathbf{u}_i$, $i = 1, 2, \ldots, n$. He shows that the decision rule to accept H_0 when $r_m^2 \le k$ and accept H_m when $r_m^2 > k$ is invariant admissible when $\gamma_i = \gamma(1 - v_{ii})^{-1/2}$, $i = 1, 2, \ldots, n$, but also admits that the rule is "not easily justifiable" when the v_{ii} are not equal. Ferguson's rule is essentially the same as one of the rules proposed by Anscombe (1960).

Although these early methods are based on somewhat different statistical philosophies, they all rely on the maximum absolute Studentized residual to furnish information about a single outlier. The work of Ferguson and Srikantan indicated that the Studentized residuals reflect the proper way to weight the ordinary residuals when the v_{ii} are nonconstant, a conclusion that was later supported by Anscombe and Tukey (1963), Andrews (1971), and Behnken and Draper (1972). Using the projection of $\mathbf{e}/(\mathbf{e}^T\mathbf{e})^{1/2}$ onto the columns of $\mathbf{I} - \mathbf{V}$, Andrews (1971) developed a test statistic for a single outlier that is statistically equivalent to r_m^2.

Despite the early work, some did not view the residuals as a useful basis for detecting outliers. Bross (1961) called Anscombe's (1960) and Daniel's (1960) work a "false start" on the grounds that the residuals are inherently unreliable since they (a) require the assumption of a specific null model, (b) are correlated, and (c) may be affected by a single outlier in complicated ways. These concerns are, of course, quite appropriate. Outliers must be judged with some model in mind. If the null model (4.1) is not a reasonable approximation in the absence of contaminants, we cannot expect to be able to accurately judge the worth of a single case. Indeed, as indicated in the Introduction, outliers should be treated generally as an indication that either the model or the cases may be in error, and they often provide useful diagnostic information. Bross's suggestion that an outlier be judged relative to a perceived pattern in the data rather than relative to an a priori specified model may occasionally provide important information, but it seems impossible to generalize in a useful way.

The correlation between the ith and jth, $i \ne j$, residuals is

$$\rho_{ij} = -v_{ij}/[(1 - v_{ii})(1 - v_{jj})]^{1/2} \qquad (4.11)$$

The importance and properties of these correlations were first studied in detail by Anscombe (1960), who provides a listing for various designs, and Anscombe (1961). Cook (1979) investigates the configurations of the rows of $\mathbf{X}$ that give rise to high correlations in regression. Cook and Prescott (1981) extend Anscombe's (1960) listing for certain models of 2^n factorials. In the presence of high correlations it may not be possible to identify the single most probable contaminant. Consider, for example, a 2×3^2 experiment with main effects and first-order interactions. For this design and model, every residual has a -1 correlation with a second residual, so that a single outlying case

cannot be identified. As suggested by Anscombe (1960), the use of the maximum absolute Studentized residual to identify a single outlying case requires the assumption that $\max |\rho_{ij}| < 1$.

Bross's final concern (c) has been expressed by others who then propose alternative criteria for the detection of a single contaminant. Mickey, Dunn, and Clark (1967), Goldsmith and Boddy (1973), and Mickey (1974) suggested that the most likely contaminant is the case that when deleted results in the largest reduction of the residual sum of squares for (4.1). Snedecor and Cochran (1967) suggested a technique based on deleting a case and comparing the associated response to its prediction obtained from fitting (4.1) to the reduced data. The fact that these techniques are statistically equivalent to r_m^2 was demonstrated by Ellenberg (1976); see also Williams (1973) and Snedecor and Cochran (1980).

In recent years r_m^2 has become a widely accepted statistic for detecting a single outlier in linear models and much work has been done on the problem of determining associated critical and p values. The null distribution of r_m^2 has been investigated by Hartley and Gentle (1975), but the results are quite complicated. One of the difficulties is that the distribution of r_m^2 depends on $\mathbf{X}$ so that a tabulation of exact percentage points is not practical in general. Monte Carlo studies that reflect the dependence of the distribution of r_m^2 on $\mathbf{X}$ are reported by Tietjen, Moore, and Beckman (1973), Ellenberg (1976), and Gentle (1978). Most of the recent work on this problem involves the use of Bonferroni bounds to determine approximate critical and p values.

The approximation of $\Pr(r_m^2 \geq d)$ using the first Bonferroni upper bound will be exact if

$$1 + \max_{i<j} |\rho_{ij}| < 2(n-p)d^2. \tag{4.12}$$

This is equivalent to the sufficient conditions given by Srikantan (1961), Stefansky (1971,1972), and Prescott (1977); see also Cook and Prescott (1981) and Ellenberg (1976).

Stefansky (1971,1972) presents a method, based on Bonferroni bounds, for calculating critical values of the maximum normal residual

$$\text{MNR} = \max |e_i| / (\mathbf{e}^T\mathbf{e})^{1/2} \tag{4.13}$$

when the v_{ii} are constant, and provides selected critical values for two-way tables. The MNR is, of course, equivalent to r_m^2 when the v_{ii} are equal. Although not easily implemented, Stefansky's method can be used to tabulate approximate critical values for experimental design models in which $\mathbf{X}$ is constant across applications. Galpin and Hawkins (1981) use Stefansky's basic approach to tabulate critical values for two- and three-way tables.

Using Monte Carlo methods, Tietjen, Moore, and Beckman (1973) investigate the dependence of the critical values of r_m^2 on the arrangement of the x's in simple linear regression. They conclude that this dependence is not strong and present critical values of $|r_m|$ obtained by averaging over four distinct arrangements. Prescott (1975) noted that the critical values given by Tietjen, Moore, and Beckman are quite close to those obtained using the first Bonferroni bound and suggested that this bound would usually produce adequate critical values for multiple regression. Following Prescott's suggestion, Lund (1975) constructed tables of approximate critical values. Moses (1978) provides charts for finding the upper percentage points of Student's t in the range .01 to .00001, and Bailey (1977) provides tables of the Bonferroni t statistic. John and Prescott (1975) compare a variety of schemes for obtaining critical values for r_m^2 and conclude that the first Bonferroni bound gives the most satisfactory result.

Ellenberg (1973,1976) suggests that the accuracy of the first Bonferroni bound might generally be checked using upper and lower bounds based on the first- and second-order Bonferroni inequalities. Evaluation of this lower bound, however, requires numerical integration and the residual correlations, and is therefore not well suited for routine use. Cook and Prescott (1981) present a relatively simple method for determining upper and lower bounds on p-values associated with r_m^2. This method explicitly incorporates the condition (4.12) necessary for the Bonferroni upper bound to be exact and allows a more accurate reflection of the weight of evidence against the null hypothesis of no outliers than is possible using the upper bound alone. Doornbos (1981) used a similar approach to construct bounds for $\max_{i<j} |\rho_{ij}|$ that when satisfied guarantee a test size between α and $(\alpha - \alpha^2/2)$.

In summary, the maximum absolute Studentized residual $|r_m|$ has been rediscovered numerous times over the past 20 years and is now a widely accepted criterion for detecting a single outlier in linear regression. Critical values or, preferably, p values based on the first-order Bonferroni upper bound can be expected to be reasonably accurate whenever the residual correlations are not excessively large. If special tables of known accuracy are lacking, the methods of Cook and Prescott (1981) or Doornbos (1981) can be used to assess the accuracy of this bound.

Unlabeled Model, $k > 1$. When one is faced with the possibility of multiple contaminants, the problem is how to make inferences about $\mathbf{i}$ and $\boldsymbol{\gamma}$ simultaneously. As in the case of simple random samples, one is also faced with the specification of the exact number of outliers, k, or an upper bound on the number of

outliers, k_u. Mickey, Dunn, and Clark (1967) avoid the choice of k_u by employing a forward selection procedure based on successively deleting individual cases so that each deletion achieves the maximum decrease in the residual sum of squares. The usual type of forward selection rules are used to indicate when to stop. Gentle (1978) proposes a related procedure, based on using $|r_m|$ stepwise to test for the presence of a single contaminant in various reduced data sets. Masking will likely present substantial problems in any procedure beginning with $k = 1$.

Gentleman and Wilk (1975a) study the properties of the Studentized residuals from analyses of two-way tables based on additive models when one or two contaminants are present. They conclude that the residuals are reliable indicators of a single outlier but that when two contaminants are present "the resulting residuals will often not have any noticeable peculiarities" (see also Daniel 1978 and Cook 1979). They later (Gentleman and Wilk 1975b) recommend that, in general, the k most outlying cases be determined by first choosing k_u and then finding the maximum value of $Q_k(\mathbf{i})$ (see eq. 4.6) for $k = k_u$. If $\max_{\mathbf{i}} Q_{k_u}(\mathbf{i})$ is not statistically discrepant the procedure is repeated with $k = k_u - 1$; otherwise the process stops with the identification of $\hat{\mathbf{i}}$ and the value of $\mathbf{i}$ that maximizes $Q_{k_u}(\mathbf{i})$ as the indices of the outliers in the data. No outliers are declared when $Q_k(\mathbf{i})$ is not statistically discrepant for $k = k_u, k_{u-1}, \ldots, 1$. They suggest comparing $Q_k(\hat{\mathbf{i}})$ to simulation results obtained when no contaminants are present as an aid to assessing the discrepancy of $Q_k(\hat{\mathbf{i}})$ at each stage. Gentleman (1980) shows how the structure of $\mathbf{V}$ can be exploited to reduce the amount of necessary computation and suggests several approximate procedures when such a reduction is not sufficient or possible. However, she offers no further advice on how to judge the significance of the final results.

John and Draper (1978) give good discussions of various methods based on $Q_k(\mathbf{i})$ and suggest a two-stage procedure for detecting up to $k_u = 2$ outliers in two-way tables (see also John 1978). This procedure is extended to $k_u = 3$ in Draper and John (1980). For $k_u = 3$, the John-Draper procedure requires that $\hat{\mathbf{i}}$ and $Q_3(\hat{\mathbf{i}})$ be obtained at the outset. Then at stage 1 $Q_3(\hat{\mathbf{i}})$ is tested for significance using smoothed Monte Carlo percentage points. If the results are not significant the procedure stops and it is concluded that $k = 0$. Otherwise, it is concluded that $1 \leq k \leq 3$ and one then proceeds to stages 2 and 3 to sort out the number of contaminants indexed by $\hat{\mathbf{i}}$. The percentage points recommended for stages 2 and 3 were again obtained by smoothing Monte Carlo percentage points.

The value of the various procedures based on $Q_k(\hat{\mathbf{i}})$ for $k > 1$ depends on context. They are cumbersome, not easily interpreted, and not well suited as routine diagnostic tools. If the null model is known to be reasonably accurate in the absence of contaminants, and one's interest is in detecting unrelated tail contaminants with large values, the $Q_k(\hat{\mathbf{i}})$ criterion may be useful. Simulation seems necessary to obtain guidance in the interpretation unless the analysis is of a two-way table and $k_u \leq 3$. (The first Bonferroni upper bound will likely be very conservative.)

As an alternative to using $Q_k(\mathbf{i})$, Bradu and Hawkins (1982) suggest a procedure, based on tetrads, for identifying multiple outliers in two-way tables.

Dempster and Gasko-Green (1981) suggest an alternative point of view based on sequentially removing the single most discrepant case from a least squares analysis. They consider various measures of discrepancy, including $|r_m|$, and argue that data analysts should consider applying more than one measure to a given set of data. As a means of judging how far the sequential deletion procedure should be carried out, they propose two sequences of conditional p values that can also be used to identify groups of outlying cases. This methodology has definite advantages over that based on $Q_k(\mathbf{i})$: It is easier to implement, produces results that are easier to interpret, and is more suitable for use as a routine diagnostic. Although this methodology is promising and seems to work well in the examples considered by Dempster and Gasko-Green, more experience is necessary before we can be definitely convinced of its value. In particular, a comparison with the methodology based on $Q_k(\mathbf{i})$ could be most interesting.

4.3 Other Approaches

Numerous approaches that do not fall within the framework of Section 4.1 have been proposed for dealing with outliers in normal linear models. These approaches differ with respect to the amount of a priori information available on the nature and frequency of the outliers, the philosophical basis for the underlying analysis and the treatment of outliers, and ease of implementation. The purpose of the section is to illustrate the kinds of approaches that have been taken and describe a few methods that seem potentially useful. We do not attempt a survey of all of the available alternative methods.

A Quick Test. As was mentioned previously, the Studentized residuals are useful diagnostics for data analysis in general, and the maximum absolute Studentized residual is a reliable statistic for detecting a single contaminant. If more than one contaminant is present, however, an inspection of the Studentized residuals alone may lead to erroneous conclusions, although residual plots will occasionally reveal multiple contaminants. The sequential methods based on $Q_k(\mathbf{i})$ require special computer programs and are not well suited for routine application.

Brown and Kildea (1979) (see also Brown 1975) suggest a quick test for asymmetric contaminants. They argue that in the presence of asymmetric contamination the residuals from a least squares fit should tend to be of one sign and that tests using the signs of residuals are useful as quick indicators of whether further analysis is required. Let **S** denote an n vector with ith element sign (e_i), $i = 1, 2, \ldots, n$. The test criterion proposed by Brown and Kildea is

$$g = \mathbf{S}^T\mathbf{V}\mathbf{S}, \tag{4.14}$$

where **V** is the hat matrix as defined at (4.3). This statistic is simply the sum of squares for the regression of **S** on **X** and thus can be computed using standard regression programs. Under the hypothesis of no outliers, g is asymptotically distributed as $(1 - 2/\pi)$ times a chi-squared random variable with p degrees of freedom. The implicit alternative hypothesis is that there is at least one contaminant present. Brown and Kildea develop their test for linear models with a general density for $\boldsymbol{\varepsilon}$; the above results are for the normal case. An intuitive feeling for the conditions under which g is likely to work well can be gained by exploiting its relationship to discriminant analysis.

Robust Regression. Like the robust methods discussed in Section 3.1, robust regression methods and in particular M estimators are usually suggested for use in situations where the iid elements of $\boldsymbol{\varepsilon}$ follow a distribution F that is symmetric about zero and has tails that are heavier than those associated with the normal distribution. The rationale behind such methods is that realizations from the tails of F can have substantial influence on the least-squares estimator, so the corresponding cases should be downweighted. There has been considerable work in this area in recent years. Hogg (1979) provides a good introduction to the literature; for more details, see the recent book by Huber (1981). Here we briefly discuss robust regression as it relates to the problem of outliers.

A common paradigm for motivating the use of robust regression when contaminants are likely to be present is that the basic distribution is $N(0, \sigma^2)$, while the contaminants arise from a heavy-tailed distribution G that is symmetric about 0. If the contaminants occur independently and α is the probability of a contaminant on any given run then, in an obvious notation,

$$F = (1 - \alpha)N(0, \sigma^2) + \alpha G. \tag{4.15}$$

M estimation is commonly recommended as an omnibus method for downweighting the cases that correspond to realizations in the tails of G.

The above formulation represents a conceptual, nonparametric basis for motivating the use of robust regression rather than a specific model for estimation, since G is not usually specified. It also reflects a concern for global weaknesses in the normal model and is clearly a method of accommodation. It is also worth noting that under this formulation contaminants are modeled by modifying the error structure of (4.1), while the mean-shift model is obtained by modifying the mean structure of (4.1). As a result, and in contrast to the mean-shift model, contaminants under (4.15) do furnish information, albeit less reliable, about $\boldsymbol{\beta}$.

It is possible to construct methods, as byproducts of using M estimation to accommodate contaminants, for the identification of outliers. When robust regression is viewed as iteratively reweighted least squares (see, for example, Holland and Welsch 1977), the weight from the final iteration may be used to identify cases for further study; the cases with the smaller weights receive the most attention. Also, the residuals from a robust fit can be inspected for anomalies in much the same way as the residuals from a least squares fit. These methods have not, however, been developed to the point where significance tests are available.

Andrews (1974) and more recently Carroll (1980) give examples of the use of robust regression when contaminants are likely to be present, and they contrast their results with those obtained by other authors using different methods. Prescott (1980) provides a selective review and examples to illustrate how robust methods can be used to detect multiple outliers. When the results from robust regression are contrasted with sequential methods based on $Q_k(\mathbf{i})$, it is generally found that conclusions regarding the outliers in the data are similar and that robust regression methods are much easier to apply, thus more suitable as routine diagnostic techniques.

Alternative Models. A number of specific alternative models for dealing with outliers in normal linear models have been proposed. Some of these reflect unique circumstances and are not suitable for general application, while others have been suggested as competitors to the mean-shift model.

Elashoff (1972) described a quadratic model for the accommodation of outliers in simple linear regression. The error distribution for Elashoff's model is given by (4.15) with $G = N[c(x - x_{\min})^2, \sigma^2]$, where $x_{\min}$ is the smallest value of the explanatory variable and the parameter c, the slope β_1, the intercept β_0, and σ^2 are unknown. The mixing parameter α was assumed known on the grounds that accurate estimation is difficult and that the results are insensitive to the choice of α in any "reasonable range." Elashoff investigated the small-sample properties of the maximum likelihood estimators of β_0, β_1, c, and σ^2 but did not give specific recommendations for the identification of the contaminants.

Aitkin and Wilson (1980) suggest that an adequate treatment of outliers can be based on mixture models of the form given at (4.15) by allowing G to consist of several normal components with possibly different means and variances, $G = \sum_{i=1}^{m} \delta_i N(\mu_i, \sigma_i^2)$, where $0 \le \delta_i \le 1$ and $\sum \delta_i = 1$. They argue that the EM algorithm provides a relatively simple way to obtain maximum likelihood estimates and that outlier identification can be based on the posterior probability that the ith case comes from the jth component, $j = 0, 1, 2, \ldots, m$. To illustrate their methodology they analyze Brownlee's stack loss data with mixture models having one to five components with equal variances but unequal means. Their example clearly shows that the results can depend critically on the number m of components used in the model. Likelihood-ratio tests for the number of components may be useful, but it seems to us that this approach basically replaces the problem of choosing k in the unlabeled mean-shift model with the problem of choosing m. Choosing m will be subject to the same types of difficulties and will inevitably lead us to similar sequential procedures when m is unknown. This approach will be most useful when we know that the contaminants consistently arise from a known number of sources.

Bayesian Methods. Box and Tiao (1968) were evidently the first to consider Bayesian methods for dealing with outliers in normal linear models. Their approach is based on the assumptions that the prior for $(\boldsymbol{\beta}, \sigma)$ is proportional to σ^{-1} and that the error distribution is of the form given at (4.15) with $G = N(0, a^2\sigma^2)$ and $a^2 > 1$, although they provide many relevant details for a more general formulation in which F is a mixture of two arbitrary distributions. The mixing parameter α and a^2 are assumed fixed.

Let $p_{\mathbf{i}}(\boldsymbol{\beta} \mid \mathbf{Y})$ denote the posterior distribution of $\boldsymbol{\beta}$ given that the cases indexed by $\mathbf{i}$ are contaminants (that is, are realizations from the $N(0, a^2\sigma^2)$ component) and let $p_0(\boldsymbol{\beta} \mid \mathbf{Y})$ denote the posterior when all cases are "good." Box and Tiao show that $p_{\mathbf{i}}(\boldsymbol{\beta} \mid \mathbf{Y})$ is a p-dimensional multivariate t distribution and that the full posterior for $\boldsymbol{\beta}$ can be expressed as a weighted sum of 2^n distributions,

$$P(\boldsymbol{\beta} \mid \mathbf{Y}) = w_0\, p_0(\boldsymbol{\beta} \mid \mathbf{Y}) + \sum_{\mathbf{i}} w_{\mathbf{i}}\, p_{\mathbf{i}}(\boldsymbol{\beta} \mid \mathbf{Y}), \quad (4.16)$$

where the summation extends over all $2^n - 1$ combinations for $\mathbf{i}$ and the weights $w_{\mathbf{i}}$ are the posterior probabilities that the cases indexed by $\mathbf{i}$ are the only contaminants; w_0 is the posterior probability of no contamination. Since the contaminants do contain information about $\boldsymbol{\beta}$ in this formulation, the conditional posterior distribution $p_{\mathbf{i}}(\boldsymbol{\beta} \mid \mathbf{Y})$ is not the same as the posterior obtained by deleting the cases indexed by $\mathbf{i}$, although it is essentially the same for large values of a^2.

For use in practice, Box and Tiao suggest including only selected terms in the expression for $p(\boldsymbol{\beta} \mid \mathbf{Y})$ and renormalizing the weights: If no contaminants are suspected only the lead term in (4.16) is relevant, and $p(\boldsymbol{\beta} \mid \mathbf{Y}) = p_0(\boldsymbol{\beta} \mid \mathbf{Y})$; if at most one contaminant is suspected,

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) \propto w_0\, p_0(\boldsymbol{\beta} \mid \mathbf{Y}) + \sum_{i=1}^{n} w_i\, p_i(\boldsymbol{\beta} \mid \mathbf{Y});$$

if at most two contaminants are suspected we add the term $\sum_{i<j} w_{ij}\, p_{ij}(\boldsymbol{\beta} \mid \mathbf{Y})$, and so on. In this approach the problem of specifying the number of outliers becomes the problem of choosing the number of terms for $p(\boldsymbol{\beta} \mid \mathbf{Y})$. Box and Tiao argue that this can often be done on prior grounds by considering α and n. If, for example, $\alpha = .05$ and $n = 20$ we might ignore all terms where $\mathbf{i}$ involves three or more components, since the prior probability of more than two contaminants is small. The fact that α and a^2 need to be specified by the experimenter is annoying, but Box and Tiao suggest that some analyses will not be sensitive to this choice.

In the reduced form for $p(\boldsymbol{\beta} \mid \mathbf{Y})$ the renormalized weight can be used to indicate the most likely contaminants. For large a, the weights for at most a single contaminant are

$$w_i = \frac{c\alpha}{(1-\alpha)a}(1 - v_{ii})^{-1/2} \times (1 - r_i^2/(n-p))^{-(n-p)/2} + O(a^{-2}), \quad (4.17)$$

where c is chosen so that the weights sum to one. The important point is that the case with the highest posterior probability of being a contaminant does not in general correspond to r_m^2; see also Cook, Holschuh, and Weisberg (1982).

Continuing the work of Guttman (1973a), Guttman, Dutter, and Freeman (1978) develop an alternative to the Box-Tiao approach. Recall that in the Box-Tiao formulation the number of contaminants is a binomial (n, α) random variable and the associated errors are assumed to arise from a $N(0, a^2\sigma^2)$ distribution. In contrast, Guttman, Freeman, and Dutter condition on the number of contaminants k and adopt the mean-shift model (4.2). The marginal posterior $q(\boldsymbol{\beta} \mid \mathbf{Y}, k)$ for $\boldsymbol{\beta}$ is derived under the assumptions that the prior for the parameters $(\boldsymbol{\beta}, \gamma, \sigma)$ of the mean-shift model given at (4.2) is proportional to σ^{-1} and that all $\binom{n}{k}$ subsets of cases are equally likely to contain the k contaminants. This model is a direct Bayesian analog of the frequentist approach discussed in Section 4.1. As in the Box-Tiao approach, $q(\boldsymbol{\beta} \mid \mathbf{Y}, k)$ can be written as a convex combination of multivariate Student's t distributions,

$$q(\boldsymbol{\beta} \mid \mathbf{Y}, k) = \sum_{\mathbf{i}}{}' w_{\mathbf{i}}\, q_{\mathbf{i}}(\boldsymbol{\beta} \mid \mathbf{Y}), \quad (4.18)$$

where the summation extends over all $\binom{n}{k}$ indices containing k components, and $q_{\mathbf{i}}$ is the posterior obtained by deleting the cases indexed by $\mathbf{i}$ and assuming no contaminants in the remaining data. Thus, q is simply a convex combination of reduced data posteriors.

The weights can be written informatively as

$$w_{\mathbf{i}} \propto [\mathbf{e}^T\mathbf{e} - Q_k(\mathbf{i})]^{-(n-p-k)/2} |\mathbf{I} - \mathbf{V}_{\mathbf{i}}|^{-1/2}, \quad (4.19)$$

which, when $k = 1$, reduces to

$$w_i \propto [1 - r_i^2/(n-p)]^{-(n-p-1)/2}(1 - v_{ii})^{-1/2}. \quad (4.20)$$

The weights in (4.20) are essentially the same as those obtained in the Box-Tiao approach when a^2 is large (see eq. (4.17)). Again we see that the subset $\tilde{\mathbf{i}}$ most likely to contain the k contaminants will not in general equal the corresponding subset $\hat{\mathbf{i}}$ from the frequentist approach discussed in Section 4.1, although when n is large and $\mathbf{V}$ is well behaved the factor involving $Q_k(\mathbf{i})$ in (4.19) will dominate $w_{\mathbf{i}}$ and we can expect to have $\tilde{\mathbf{i}} = \hat{\mathbf{i}}$. In small data sets the results can be quite different.

Guttman, Dutter, and Freeman (1978) suggest basing the choice of k on the traces of the covariance matrices of $q(\boldsymbol{\beta} \mid \mathbf{Y}, k)$ for $k = 0, 1, 2, \ldots, k_u$. The smallest trace correspond to the most likely (in some sense) value of k. The value of this procedure is unclear.

Abraham and Box (1978) apply the ideas set forth by Box and Tiao to the mean-shift model, subject to the constraint that all contaminants come from the same population so that $\boldsymbol{\gamma} = \gamma\mathbf{1}$. Assuming that the prior for $(\boldsymbol{\beta}, \gamma, \sigma)$ is proportional to σ^{-1} and that the contaminants arise independently with fixed probability α, the marginal posterior is again a convex combination of multivariate Student's t distributions of the form given at (4.16). Apart from factors that do not depend on i, the posterior probability that the ith case is the only contaminant is given by (4.20).

In the approaches of Box and Tiao (1968) and Abraham and Box (1978), the contaminants do contain information about $\boldsymbol{\beta}$, in the sense that the posteriors for $\boldsymbol{\beta}$ obtained by conditioning on $\mathbf{i}$ are not the same as the reduced posterior obtained by discarding the cases indexed by $\mathbf{i}$ and assuming that the remaining cases are "good." In the approach of Guttman, Dutter, and Freeman the contaminants contain no information about $\boldsymbol{\beta}$ since the conditional and reduced posteriors are the same. Each of these approaches can be regarded as a method of accommodation based on the appropriate posterior distribution for $\boldsymbol{\beta}$, or a method of identification based on the posterior probability $w_{\mathbf{i}}$ that $\mathbf{i}$ indexes the contaminants.

There are a number of ways in which these methods might be extended. For example, the formulation by Guttman, Dutter, and Freeman could be usefully modified by allowing k to be a binomial (α, n) random variable and perhaps adopting a prior for α as well. These ideas, as well as many others, are set forth in unpublished technical reports by Bailey and Box (1980a, b), who describe the following general paradigm for dealing with outliers: Let M be a generic symbol for the model, and let $M_{\mathbf{i}}$, M_k, and M_α denote the models that are formulated conditionally on $\mathbf{i}$, k, and α, respectively. Then

$$M_{\mathbf{i}} \subset M_k \subset M_\alpha \subset M, \quad (4.21)$$

where M denotes the most general model obtained by adopting a prior for α that reflect the experimenter's attitude about the possibility of contaminants. Bailey and Box pass (a) from $M_{\mathbf{i}}$ to M_k by assigning equal prior probabilities to all $\binom{n}{k}$ subsets, (b) from M_k to M_α by assuming that k has a binomial (α, n) prior, and (c) from M_α to M by assuming that α has a beta density. The approach of Guttman, Dutter, and Freeman (1978) is in terms of M_k, while Box and Tiao (1968) and Abraham and Box (1978) work in terms of M_α. Bailey and Box (1980a) give the relevant details for methods based on M, but there are too many of these to allow discussion of them here. The general inference pattern is, however, similar to that discussed above. One important point is that, in contrast to most other suggestions, the general approach of Bailey and Box does provide a firm rational basis for inference about k.

4.4 Influential Cases

When a few outlying cases have been identified, a useful bit of practical advice that is often offered (see, for example, Kruskal 1960 and Snedecor and Cochran 1967) is to remove the cases in question and reanalyze the data. If the important aspects of the analysis are essentially unchanged, there is no reason to worry about the outliers, unless, of course, identifying cases for further study is important. On the other hand, if the analyses differ considerably, the conclusions will hinge on the assumptions, prior information, and method of analysis. In recent years more direct approaches to the study of influential cases have been taken. The results of these studies enable the investigator to obtain a better understanding of the structure of a problem in terms of the contributions of individual cases or of groups of cases than is available from the usual methods. There is now a considerable literature on influential cases and many references are given in the Introduction. Here we give a brief account of a few central developments.

The basic idea behind the study of influential cases is to monitor the changes in a selected phase of the analysis as individual cases or groups of cases are in turn removed from the data. If $\hat{\boldsymbol{\beta}}$ is of primary interest,

implementation of this idea amounts to comparing $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_{(\mathbf{i})}$ as $\mathbf{i}$ varies over the $\binom{n}{k}$ subsets of size k. In particular, it will generally be sufficient to consider the differences $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\mathbf{i})}$. A more revealing development can be based on various empirical versions of the influence curve as in Cook and Weisberg (1982), but for present purposes a heuristic approach will suffice.

One immediate difficulty is that $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\mathbf{i})}$ is a p vector and a routine inspection of $\binom{n}{k}$ such vectors is clearly not possible. A useful solution is to measure the distance $D_{\mathbf{i}}(\mathbf{M}, c)$ between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(\mathbf{i})}$ relative to a selected positive semidefinite inner-product matrix $\mathbf{M}$ and a positive scale factor c,

$$D_{\mathbf{i}}(\mathbf{M}, c) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\mathbf{i})})^T \mathbf{M} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(\mathbf{i})})/c \qquad (4.22)$$

This class of norms is used to measure the influence of the cases indexed by $\mathbf{i}$ on $\hat{\boldsymbol{\beta}}$ and not as the basis for significance tests, but see Dempster and Gasko-Green (1981). The matrix $\mathbf{M}$ and c can be chosen to reflect specific concerns. There is considerable flexibility in these choices, and most of the proposed influence measures can be written in the form of (4.22), particularly if we allow $\mathbf{M}$ and c to depend on $\mathbf{i}$. Such a dependence does not, however, seem desirable: any reasonable norm should allow the individual cases to be unambiguously ordered on the basis of influence. Such an ordering may not be possible when $\mathbf{M}$ and/or c depend on i, for the metric used to compare $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_{(\mathbf{i})}$ will vary with i. At the very least, such pseudo-orderings must be interpreted with great care. The norm suggested by Belsley, Kuh, and Welsch (1980) is equivalent to $D_i(\mathbf{X}^T\mathbf{X}, p\hat{\sigma}^2_{(i)})$ and thus the metric for this influence measure does depend on i.

A useful measure of influence for $\hat{\boldsymbol{\beta}}$ is obtained by setting $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and $c = p\hat{\sigma}^2$ (Cook 1977,1979) to yield, after some algebra,

$$D_i = D_i(\mathbf{X}^T\mathbf{X}, p\hat{\sigma}^2) = \frac{r_i^2}{p}\left(\frac{v_{ii}}{1 - v_{ii}}\right). \qquad (4.23)$$

This and related distance measures have a variety of useful interpretations, as discussed in Cook and Weisberg (1980,1982). In particular, when D_i is compared to the percentage points of an F distribution with p and $n - p$ degrees of freedom, the corresponding "p value" gives the level of the smallest confidence ellipsoid based on the full data that contains $\hat{\boldsymbol{\beta}}_{(i)}$. Any two cases i and j for which $\boldsymbol{\beta}_{(i)}$ and $\boldsymbol{\beta}_{(j)}$ falling on the edge of the same full-data confidence ellipsoid will thus be judged equally influential. If one desires, a robust estimator of σ^2 can be used in place of $\hat{\sigma}^2$.

From expression (4.23) it is clear that outlying cases (r_i^2 large) need not be influential if v_{ii} is sufficiently small. This can happen, for example, if the corresponding $\mathbf{x}_i$ lies near the average $\bar{\mathbf{x}}$ of the rows of $\mathbf{X}$. Conversely, a nonoutlying case may be highly influential if v_{ii} is sufficiently large (the corresponding $\mathbf{x}_i$ lies far from $\bar{\mathbf{x}}$). This fundamental interaction between r_i^2 and v_{ii} is characteristic of all of the ways that have been proposed for measuring the influence of a single case on $\hat{\boldsymbol{\beta}}$ and, interestingly, is also characteristic of the Bayesian approaches to outlier identification (see eqs. (4.17) and (4.20)) but not of the frequentist approaches.

The notion of confidence ellipsoid displacement that characterizes D_i can be used to construct a useful influence measure for $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$. The idea is to use the contours of the full-sample log-likelihood to measure the distance between $(\boldsymbol{\beta}, \hat{\sigma}^2)$ and $(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\sigma}^2_{(i)})$. The resulting likelihood distance is (see Cook and Weisberg 1982, for details)

$$LD_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = n \log\left[\left(\frac{n}{n-1}\right)\frac{n-p-1}{F_1(i) + n - p - 1}\right] + \frac{(n-1)F_1(i)}{(1 - v_{ii})(n - p - 1)} - 1, \qquad (4.24)$$

where $F_1(i)$ is defined at (4.10). This measure can be compared to the chi-squared distribution with $p + 1$ degrees of freedom for calibration, just as D_i is compared to the F distribution. The corresponding likelihood distance for $\hat{\boldsymbol{\beta}}$ alone is $LD_i(\hat{\boldsymbol{\beta}}) = n \log [pD_i/(n - p) + 1]$, which is monotonic in, and thus equivalent to, D_i.

Andrews and Pregibon (1978) suggested using the ratio

$$R_{\mathbf{i}} = \frac{(n - p - k)\hat{\sigma}^2_{(\mathbf{i})} |\mathbf{X}^T_{(\mathbf{i})}\mathbf{X}_{(\mathbf{i})}|}{(n - p)\hat{\sigma}^2 |\mathbf{X}^T\mathbf{X}|} \qquad (4.25)$$

for identifying subsets of k influential cases. Here, $\mathbf{X}_{(\mathbf{i})}$ is obtained by deleting the k rows indexed by $\mathbf{i}$ from $\mathbf{X}$ and $\sigma^2_{(\mathbf{i})}$ is defined following (4.5). The rationale for this measure is based on the idea that the deletion of an outlier in $\mathbf{Y}$ will result in a marked reduction in the residual sum of squares and the deletion of a remote row of $\mathbf{X}$ will produce a similar change in $|\mathbf{X}^T\mathbf{X}|$. Thus, small values of $R_{\mathbf{i}}$ correspond to influential subsets.

Let $\mathbf{X}^* = (\mathbf{X}, \mathbf{Y})$ be the matrix of explanatory variables augmented with $\mathbf{Y}$. Then (Draper and John 1981)

$$R_{\mathbf{i}} = |\mathbf{X}^{*T}_{(\mathbf{i})}\mathbf{X}^*_{(\mathbf{i})}| / |\mathbf{X}^{*T}\mathbf{X}^*| \qquad (4.26)$$

$$= (1 - (Q_k(\mathbf{i})/\mathbf{e}^T\mathbf{e}))\,|\mathbf{I} - \mathbf{V}_i| \qquad (4.27)$$

and, when $k = 1$,

$$= (1 - (r_i^2/(n - p)))(1 - v_{ii}). \qquad (4.28)$$

Expression (4.26) shows that $R_{\mathbf{i}}$ does not give special attention to the response vector $\mathbf{Y}$ and thus does not exploit the full structure of the linear model. Expressions (4.27) and (4.28) are provided for comparison with D_i in (4.23) and the corresponding Bayesian statistics (4.17), (4.19), and (4.20). Draper and John

(1981) examine the relative merits of R_i and D_i. They show that a case that is influential based on R_i need not be so according to D_i and vice versa, and recommend D_i, $Q_k(i)$, and $|\mathbf{I} - \mathbf{V}_i|$ for routine use.

4.5 Remarks

There is clearly a somewhat confusing array of methods for dealing with outlying cases in normal linear models. Our experience with these methods leads to the rather weak conclusion that most of the time most of the methods will yield essentially the same results. This gives some indication that we may be on the right track, but it is of little help to investigators faced with their own unique problems.

The situation when $k = 1$ is the most clearcut: If prior knowledge to indicate that the possible contaminants carry information about $(\boldsymbol{\beta}, \sigma^2)$ is lacking, analyses should be based on r_i^2, v_{ii}, and some combinations thereof that measure the influence of the ith case on important phases of the analysis. The r_i^2 result from the mean-shift model, and this is essentially the only model that treats the contaminants as totally uninformative. An inspection of the v_{ii}s is important because they provide useful information about power and the relationship between $\mathbf{x}_i^T$ and the remaining rows of $\mathbf{X}$. If some of the v_{ii} are large, the decision to accept the corresponding cases may be based on faith alone since the power at such cases will be relatively small. The reasons for studying influence should be evident. The fact that an influential case is not outlying should not decrease interest in that case. In the absence of special considerations, D_i and $LD_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ are preferable measures of influence.

Generally, the above methods are best regarded as useful routine diagnostics rather than as methods based on the firm conviction that $k_u = 1$, since such a conviction will almost always be lacking.

The situation for $k > 1$ is similar provided that a specific value for k or k_u can be selected on firm, well-grounded information. The analogous quantities are $Q_k(i)$, $(\mathbf{I} - \mathbf{V}_i)$ and D_i or LD_i. Simulation might be used to determine percentage points, provided that the importance of the problem justifies the effort. However, the methods for $k > 1$ require an overwhelming commitment to the study of outliers. In our experience, the urge to "always consider the possibility of one more outlier" is often overwhelming. Indeed, the choice of k_u is usually based on computational considerations rather than on a reasonable belief that the remaining data are "clean." An easily implemented, reliable method that indicates when a deeper analysis $(k > 1)$ is required is surely needed.

The quantities $Q_k(i)$ and $\mathbf{V}_i$ are the building blocks for many of the methods discussed in this section. Computational details for these as well as many other relevant quantities can be found in Cook and Weisberg (1982) and Gentlemen (1980).

5. OTHER PROBLEM AREAS

5.1 Circular Data

For circular data, the observation most likely to be a contaminant is characterized by having the largest angular deviation from the average direction. Collett (1980) describes and compares four possible tests for a single contaminant. Two of these tests require an underlying von Mises distribution while the remaining two are based on ad hoc considerations. An investigation of the relative performance of the various tests is based on sampling from a von Mises distribution with a "mean-shift" alternative. Collett concludes that when the sample size exceeds 15 and the concentration parameter of the von Mises distribution is large, the four tests are essentially equivalent.

5.2 Discriminant Analysis

Campbell (1978) suggests using the influence function as a basis for outlier detection in discriminant analysis. He derives the theoretical and sample influence function for various statistical summaries, including Mahalanobis D^2 and the discriminant means and coefficients. Outlier detection is based on graphical displays of the sample influence functions.

5.3 Experimental Design

Box and Draper (1975) consider designs that minimize the effects of outliers on the fitted values $\hat{\mathbf{Y}}$ obtained by least squares estimation based on model (4.1). They show that $\hat{\mathbf{Y}}$ will be relatively insensitive to outliers if the v_{ii} are as uniform as possible and suggest $\sum v_{ii}^2$ as one convenient measure of uniformity. In adopting this approach, it is assumed that ordinary least squares will be used regardless of the resulting data. For example, the possibility that gross contaminants will be identified and rejected is not included as part of the criterion.

Herzberg and Andrews (1976) describe a design criterion that may be useful when outliers are identified and discarded. Let $\boldsymbol{\Delta}$ be an $n \times n$ diagonal matrix with ith diagonal element $\delta_i = 0$ if the ith case is an outlier and 1 otherwise. The δ_i are assumed to be independent random variables with $\Pr(\delta_i = 0) = \alpha(\mathbf{x}_i)$. For given $\alpha(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$, the probability that (4.1) will be rank-deficient is $\Pr(|\mathbf{X}^T\boldsymbol{\Delta}\mathbf{X}| = 0)$ and the expected precision is $E(|\mathbf{X}^T\boldsymbol{\Delta}\mathbf{X}|^{1/p})$. Designs based on these criteria will be relatively insensitive to the removal of outlying cases.

Draper and Herzberg (1979) use the minimum average integrated mean squared error as a design criterion to minimize bias resulting from the presence of outliers. This approach is basically an adaptation of

the fundamental ideas of Box and Draper (1959). It is assumed that model (4.1) will be the fitted model while the true model is some version of the mean-shift model (4.2). Draper and Herzberg derive expressions for the integrated mean squared error and examine a few typical situations in detail.

For further results on designing in the presence of outliers, see Nachtsheim (1979).

5.4 Multivariate Outliers

Techniques have been proposed to detect outliers in multivariate normal samples that are straightforward generalizations of the univariate techniques discussed in Section 3. For the detection of a single multivariate outlier, Siotani (1959) proposed the use of the generalized standardized residuals

$$X^2_{\max} = \max\,(\mathbf{X}_i - \bar{\mathbf{X}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})$$

and

$$T^2_{\max} = \max\,(\mathbf{X}_i - \bar{\mathbf{X}})^T\mathbf{S}_v^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}).$$

where $\boldsymbol{\Sigma}$ is the known covariance matrix and $\mathbf{S}_v$ is an independent estimate of $\boldsymbol{\Sigma}$. To test for a specified number k of observations as outliers, Wilks (1963) used the ratio of the generalized distance with k observations deleted to the generalized distance for all the data. Letting $\mathbf{S}$ be the sample covariance matrix and $\mathbf{S(i)}$ be that obtained by deleting the k observations indexed by $\mathbf{i}$, the Wilks statistic is obtained by taking the minimum of the $\binom{n}{k}$ possible values of $R(\mathbf{i}) = |\mathbf{S(i)}|/|\mathbf{S}|$. Tables of critical values were given by Wilks for one and two outliers.

The statistics suggested by Siotani and Wilks are multivariate generalizations of work of Thompson (1935), Pearson and Chandra Sekar (1936), Nair (1948), and Grubbs (1950). A multivariate extension of Dixon's (1950) gap test was given by Rohlf (1975). He used as a test statistic the lengths of the $n - 1$ edges of the minimal spanning tree formed from data standardized by robust estimates of the standard deviation. The lengths of these edges were shown by empirical investigation to have an approximate gamma distribution.

Guttman's (1973a) Bayesian technique for the detection of univariate outliers was extended in the same paper to multivariate samples. As in the univariate case, the posterior mean was shown to be a function of weights that indicate outlying observations.

Recently, Schwager and Margolin (1982) obtained the locally best invariant test for the multivariate mean-shift outlier model. The test procedure is an extension of Ferguson's (1961) univariate test and uses Mardia's multivariate sample kurtosis as a test statistic.

Two methods that have no univariate counterpart have been proposed for the detection of multivariate outliers. Andrews (1972) proposed plotting, for each data point $\mathbf{X}^T = (X_1, X_2, \ldots, X_m)$, the function $f(t) = X_1\sqrt{2} + X_2 \sin(t) + X_3 \cos(t) + X_4 \sin(2t) + X_5 \cos(2t) + \ldots$ over the range $-\pi < t < \pi$. These plots are used both for the detection of outliers and for cluster analysis. Hawkins (1974) suggested that principal components be used to detect errors in multivariate data. Under the assumption that an independent estimate of the covariance matrix is available, he computed the transformation $\mathbf{Z} = \mathbf{FX}$, where $\mathbf{F}$ is the matrix of normalized principal-component weights. A particular data vector $\mathbf{X}$, which is independent of $\mathbf{S}_v$, is judged as outlying if selected elements of $\mathbf{Z}$ are sufficiently large.

Many of the above procedures were suggested or reviewed by Gnanadesikan and Kettenring (1972). They gave heuristic arguments for the use of probability plots and principal-components analysis for the detection of multivariate outliers. They also encouraged the use of many test procedures for the discovery of multivariate outliers since "the complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure."

The development of large computers has certainly facilitated the analysis of multivariate data. We expect that some of the most fruitful work in the area of outliers will be in the multivariate setting. Procedures that are free from masking and swamping for testing multivariate outliers need to be developed.

For robust estimation in the multivariate cases, see Maronna (1976) and Campbell (1980).

5.5 Generalized Linear Models

Let $Y_1, \ldots, Y_n$ be n independent random variables such that Y_i has density or probability function

$$f_i(y \mid \mathbf{x}_i^T\boldsymbol{\beta}), \qquad i = 1, 2, \ldots, n, \tag{5.1}$$

where $\mathbf{x}_i^T$ is a known p-vector of explanatory variables, $i = 1, 2, \ldots, n$, and $\boldsymbol{\beta}$ is an unknown parameter vector. For exponential families, the usual problems of estimation and testing in this class of generalized linear models have been considered by Nelder and Wedderburn (1972) and Wedderburn (1976), but very little has been done on the problem of outliers. One difficulty is that the errors are not, in general, additive. One way to proceed, however, is based on the argument that any y_i can be regarded as a realization from $f_i(y \mid \mathbf{x}_i^T\boldsymbol{\beta})$ for some $\mathbf{x}_i$, so that the effective cause of a contaminant is an error in $\boldsymbol{x}_i$. The labeled model that includes the ith case as a possible contaminant can now be represented as (see the discussion following (4.2))

$$f_j(y \mid \mathbf{x}_j^T\boldsymbol{\beta} + \gamma u_j), \qquad j = 1, 2, \ldots, n, \tag{5.2}$$

where $u_j = 1$ if $j = i$ and $u_j = 0$ otherwise. The maxi-

mum likelihood estimator $\tilde{\boldsymbol{\beta}}_{(i)}$ of $\boldsymbol{\beta}$ under (5.2) is the same as that obtained from (5.1) after deleting the ith case; the ML estimator of γ is then determined as the solution to

$$l_i(x_i^T\tilde{\boldsymbol{\beta}}_{(i)} + \gamma) = 0, \tag{5.3}$$

where $l_i = (\partial/\partial a_i) \log f_i(y \mid a_i)$. Tests of the hypothesis $\gamma = 0$ can be constructed using the usual likelihood arguments, although simulation may be needed to determine percentage points since the corresponding large-sample results may not be sufficiently accurate.

The unlabeled version of (5.3) can be approached in the familiar way, but the need to determine $n + 1$ ML estimators of $\boldsymbol{\beta}$, each of which may require iteration, overshadows the usefulness of this approach. One alternative is to base inference on the maximum of the score test statistics (see Cox and Hinkley 1974, p. 324) over the n possible labeled models. The advantage of the score statistic is that it can be computed from a fit of the null model alone. This procedure should serve well for the identification of outliers, but, again, simulation may be necessary to obtain critical values.

Logistic regression is an important special case that falls within this framework. For the logistic model, $f_j(y \mid \mathbf{x}_j^T\boldsymbol{\beta})$ is the binomial $B(n_j, \alpha_j)$ probability function with $\log [\alpha_j/(1 - \alpha_j)] = \mathbf{x}_j^T\boldsymbol{\beta}$. The derivative of the log-likelihood based on the ith case is

$$l_i = y_i - n_i\alpha_i.$$

Let $\mathbf{W}$ denote an $n \times n$ diagonal matrix with ith diagonal element $-(\partial^2/\partial a_i^2) \log f_i(y \mid a_i) = n_i\alpha_i(1 - \alpha_i)$ and let $\mathbf{Z} = (z_{ij}) = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$, where $\mathbf{X}$ is the $n \times p$ matrix with rows $\mathbf{x}_i^T$, $i = 1, 2, \ldots, n$. Then the score test statistic for the hypothesis that $\gamma = 0$ in the labeled model (5.2) is

$$\tilde{l}_i^2/(\tilde{w}_{ii}(1 - \tilde{z}_{ii})), \tag{5.4}$$

where the added tilde accent indicates that all quantities are evaluated at the ML estimates from the null model. In the unlabeled model, the most outlying case is determined by maximizing (5.4) over i.

For more general discussions of logistic regression diagnostics, including various extensions of the influence measures discussed in Section 4.3, see Pregibon (1979, 1981), Cook and Weisberg (1982), and Jennings (1982).

5.6 Generalized Residuals

The ideas outlined in the previous subsection constitute a useful framework for the study of outliers when the density of y_i can be parameterized in terms of $\mathbf{x}_i^T\boldsymbol{\beta}$ alone. However, several important problems—nonlinear regression, for example—are excluded. A second general framework for outlier studies can be based on the general definition of a residual developed by Cox and Snell (1968, 1971).

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent, continuous random variables with completely known densities and assume that the ith response y_i can be represented as

$$y_i = g_i(\boldsymbol{\theta}, \varepsilon_i), \qquad i = 1, 2, \ldots, n, \tag{5.5}$$

where the g_i's are known functions that may depend on the known values of explanatory variables. For the usual linear model, for example, $g_i(\boldsymbol{\theta}, \varepsilon_i) = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma\varepsilon_i$, where $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \sigma)$ and the ε_i's are iid $N(0, 1)$. Assuming a unique solution for ε_i, (5.5) can be rewritten as

$$\varepsilon_i = h_i(\boldsymbol{\theta}, y_i), \qquad i = 1, 2, \ldots, n. \tag{5.6}$$

Cox and Snell define the ith crude residual R_i as

$$R_i = h_i(\tilde{\boldsymbol{\theta}}, y_i),$$

where $\tilde{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$.

In large samples, the R_i's should exhibit roughly the same behavior as the ε_i's and can thus be used for diagnostic purposes. In particular, "large" values of R_i, or preferably Studentized versions thereof, serve to indicate cases for further study. This technique is best viewed as an omnibus method for outlier identification, since the results will not necessarily be the same as those based on a specific alternative model.

5.7 Nonnormal Distributions

Although discordancy tests for outlying observations have been developed for almost all "ordinary" distributions, the bulk of the nonnormal literature pertains to the exponential and gamma distributions. Hence, we will limit our discussion to these distributions. The reader is referred to Barnett and Lewis (1978) for tests involving other distributions.

Letting $T(x)$ be the total life time in the interval $[0, x]$ from ordered exponential samples $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, Epstein (1960a, b) showed that the statistic $R = (n - 1)T(X_{(1)})/T(X_{(n)} - X_{(1)})$ has an $F(2, 2n - 2)$ distribution and can be used to test the smallest observation as outlying. Similarly, the smallest two observations can be tested by using the statistic $(n - 2)T(X_{(2)})/2T(X_{(n)} - X_{(2)})$.

Laurent (1963) derived the distribution of $(\bar{X} - X_{(1)})/(X_{(n)} - X_{(1)})$ for exponential samples. Basu (1965) applied this statistic in the rejection of a single observation in an exponential sample. Likés (1966) adapted Laurent's work in the derivation of Dixon's ratio-of-gaps statistics for exponential distributions. Kabe (1970) gave an algorithm for the efficient computation of Likés's critical values. Mount and Kale (1973) viewed outlier detection as a multiple-decision process for stochastically ordered distributions, the exponential being one special case. For monotone likelihood functions they showed that $X_{(1)}$ or $X_{(n)}$ are the appropriate test statistics. The likelihood ratio test $T = (X_{(n-1)} + X_{(n)})/n\bar{X}$ for two outlying exponential

observations is investigated by Kimber and Stevens (1981). It compares favorably with the Shapiro and Wilk (1972) W test for exponentiality when there are exactly two outliers.

Using L estimators, Kale and Sinha (1971) gave an estimator of the exponential mean that is robust against one outlier. Sinha (1972) gave an estimator of the reliability for an exponential distribution that is also robust against a single outlier.

By assuming that there are exactly k exponential outliers and a prior distribution on the unknown parameters, Kale (1976) showed that the most likely set of outliers in an exponential sample is the set of the k largest order statistics. He also proposed that k be determined by plots of the order statistics against their expected values as estimated by leaving out the k largest observations.

Lingappaiah (1976) assumed k outliers from the generalized density

$$f(x) = bx^{\alpha-1}\beta^{\alpha/b} \exp(-\beta x^b)/\Gamma(\alpha/b),$$

which includes the gamma, exponential, and Weibull distributions. Under the assumption that the outliers came from the above density with β replaced by $\theta_i\beta$, he found the posterior mean of β given an exponential prior on β and beta priors on θ_i.

Let $T_{(j)} = X_{(j)}/n\bar{X}$. For gamma samples, Lewis and Fieller (1979) found the distribution of $T_{(1)}$ and $T_{(n)}$ and the joint distribution of $T_{(1)}$, $T_{(2)}$ and $T_{(n-1)}$, $T_{(n)}$. In addition they give the distribution of $(X_{(2)} - X_{(1)})/(X_{(n)} - X_{(1)})$. In an earlier paper Hawkins (1972) derived the distribution function of $T_{(n)}$ for the special case of the chi-squared distribution, and investigated the power of the associated test for a single outlier.

Kimber (1979), using the suggestion of Barnett and Lewis (1978, p. 88), showed that gamma samples could be transformed to approximate normality by taking the third root of the data and investigated the use of the transformed data in the maximum Studentized residual and the sample-kurtosis test statistics. Based on likelihood arguments, he also proposed the statistic

$$Z = \{ny - (n-1)\min y_i\} \Big/ \left(n^2 y - (n-1)\sum_{i=1}^{n} y_i\right),$$

where y is the difference in the logarithm of the arithmetic mean and the geometric mean, and y_i is the same difference with the ith observation deleted. Comparing these three statistics for samples from a $\Gamma(10, 1)$ and outliers from a $\Gamma(10, s)$, Kimber found Z to be more powerful than the Studentized residual for $s < 1$ but less powerful for $s > 1$.

5.8 Time Series

Fox (1972) was evidently the first to publish a paper considering outliers in time series problems. The approach is basically an adaptation of the mean-shift model that takes into account the special structure of the problem. The process is assumed to be stationary with a known order.

Fox distinguishes between two types of outliers: Type I consists of an error that affects only a single observation, as in the mean-shift model for regression, while Type II consists of an error that affects a particular observation and all subsequent observations in the series. For the Type I outlier, Fox derives a likelihood-ratio statistic for testing that the discrepancy is zero in the labeled version of the model and uses the maximum of this statistic for the unlabeled model. Simulation is used to obtain percentage points in the latter case. For a Type II outlier only the labeled model is considered.

Abraham and Box (1979) develop a Bayesian approach to the analysis of outliers in time series. They refer to models containing Fox's Type I and Type II outliers as the aberrant-observation and aberrant-innovation models, respectively. Their basic approach is similar to those outlined in Section 4.2.

6. DISCUSSION

When we began this article, it was suggested that we separate "the wheat from the chaff" and provide a few definite recommendations for dealing with outliers. The choice of a method is not easy and depends on one's philosophy, the available a priori information (in both the Bayesian and the frequentist approaches), whether accommodation or identification is the major concern, and, of course, the specific goals of the analysis. The truth is that many of the methods that we have mentioned, and all of the methods that we have discussed in some detail, can have a place, depending on the investigator's requirements.

Should we routinely subject all samples to some type of outlier procedure? Yes. In linear models, for example, an inspection of the Studentized residuals, the diagonal elements of $\mathbf{V}$, and a single-case measure of influence should be included as a routine part of the diagnostic phase of every analysis. For simple, normal samples a visual inspection of the data using, for example, normal probability plots will often be sufficient, but recall the warnings of Collett and Lewis (1976). Generally, application of such diagnostic methods will often be valuable for guiding the subsequent analysis and may indicate the need for accommodation through transformations or other revisions of the model, or the need for application of a method of robust estimation.

Should we routinely seek to reject outliers on the basis of formal tests? No. During the model-building phase of an analysis, for example, formal tests for outliers should not be used, since they require that the null model be accurate in terms of both the expecta-

tion and the distribution of errors. Formal tests should be used only for situations in which the null model is known to be accurate in the absence of contaminants and interest centers on the study of alternative phenomena. Little information is available on the performance of the usual least squares estimators in combination with rejection via formal tests.

In normal linear models perhaps the most annoying problem that we have not been able to resolve to our satisfaction is how to deal with the possibility of multiple outliers when there is little relevant a priori information on the number and type of outliers, and identification is the major concern. Some recommended methods of identification arise as by-products of methods of accommodation. The weights from robust regression and the Bayesian methods, for example, fall into this category. Methods of accommodation, however, do rely on specific a priori information (symmetry in robust estimation, for example) and are developed with the properties of the resulting estimators in mind. It is not clear, at least not to us, that such formulations adequately meet the requirements for identification. In robust estimation, it is commonly recommended that the tuning constants be chosen to achieve 95 percent efficiency at the normal model. Is such a choice appropriate for identification or should a lower efficiency be chosen? Of course, we may try various values of the tuning constants and proceed in an exploratory manner. The tuning constants then play much the same role as an upper bound k_u on the number of outliers, and any advantages of robust estimation for the purposes of identification are uncertain.

ACKNOWLEDGMENTS

We are grateful to Christopher Bingham, Maurice Bryson, Mark Johnson, Mike McKay, Rick Picard, Sandy Weisberg, and the editorial staff for their comments on an early revision of this manuscript. Special thanks are extended to Dixie Hanks for editorial assistance, D. Whiteman, and S. Juan.

[*Received July 1982. Revised September 1982.*]

REFERENCES

ABRAHAM, B., and BOX, G. E. P. (1978), "Linear Models and Spurious Observations," *Applied Statistics*, 27, 131–138.

——— (1979), "Bayesian Analysis of Some Outlier Problems in Time Series," *Biometrika*, 66, 229–236.

AITKIN, M., and WILSON, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325–332.

AIRY, G. B. (1856), "Letter From Professor Airy, Astronomer Royal, to the Editor," *Astronomical Journal*, 90, 137–138.

ANDREWS, D. F. (1971), "Significance Tests Based on Residuals," *Biometrika*, 58, 139–148.

——— (1972), "Plots of High-Dimensional Data," *Biometrics*, 28, 125–136.

——— (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523–531.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., and TUKEY, J. W. (1972), *Robust Estimates of Location*, Princeton, N.J.: Princeton University Press.

ANDREWS, D. F., and PREGIBON, D. (1978), "Finding the Outliers That Matter," *Journal of the Royal Statistical Society*, Ser. B, 40, 87–93.

ANSCOMBE, F. J. (1960), "Rejection of Outliers," *Technometrics*, 2, 123–147.

——— (1961), "Examination of Residuals," *Proceedings of The Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press: Berkeley and Los Angeles, 1–35.

ANSCOMBE, F. J., and TUKEY, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141–159.

ATKINSON, A. C. (1981), "Robustness, Transformations and Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika*, 68, 13–20.

——— (1982), "Regression Diagnostics, Transformations and Constructed Variables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 44, 1–35.

BAILEY, B. (1977), "Tables of the Bonferroni *t*-Statistic," *Journal of the American Statistical Association*, 72, 469–478.

BAILEY, S. P., and BOX, G. E. P. (1980a), "Modeling the Nature of Frequency of Outliers," Technical Report 2085, Mathematics Research Center, University of Wisconsin at Madison.

——— (1980b), "The Duality of Diagnostic Checking and Robustification in Model Building: Some Considerations and Examples," Technical Report 2086, Mathematics Research Center, University of Wisconsin at Madison.

BARNETT, V. (1978), "The Study of Outliers: Purpose and Model," *Applied Statistics*, 27, 242–250.

BARNETT, V., and LEWIS, T. (1978), *Outliers in Statistical Data*, New York: John Wiley.

BASU, A. P. (1965), "On Some Tests of Hypotheses Relating to the Exponential Distribution When Some Outliers Are Present," *Journal of the American Statistical Association*, 60, 548–559.

BECKMAN, R. J., and TRUSSELL, H. J. (1974), "The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regression," *Journal of the American Statistical Association*, 69, 199–201.

BEHNKEN, D. W., and DRAPER, N. R. (1972), "Residuals and Their Variance Patterns," *Technometrics*, 14, 101–111.

BELSLEY, D. A., KUH, E., and WELSCH, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

BERNOULLI, D. (1777), "The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of the Most Likely Induction," in C. G. Allen (1961), *Biometrika*, 48, 3–13.

BICKEL, P. J. (1965), "On Some Robust Estimates of Location," *Annals of Mathematical Statistics*, 36, 847–858.

BICKEL, P. J., and HODGES, J. L. JR. (1967), "The Asymptotic Theory of Galton's Test and a Related Simple Estimate of Location," *Annals of Mathematical Statistics*, 38, 73–89.

BOX, G. E. P. (1979), "Strategy of Scientific Model Building," in *Robustness in Statistics*, eds. R. L. Launer and G. N. Wilkinson, New York, Academic Press

——— (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

BOX, G. E. P., and DRAPER, N. R. (1959), "A Basis for the Selection of Response Surface Design," *Journal of the American Statistical Association*, 54, 622–654.

——— (1975), "Robust Designs," *Biometrika*, 62, 347–352.

BOX, G. E. P., and TIAO, G. C. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika*, 55, 119–129.

BRADU, D., and HAWKINS, D. M. (1982), "Location of Multiple Outliers in Two-Way Tables, Using Tetrads," *Technometrics*, 24, 103–108.

BROSS, I. D. J. (1961), "Outliers in Patterned Experiments: A Strategic Appraisal," *Technometrics*, 3, 91–102.

BROWN, B. M. (1975), "A Short-Cut Test for Outliers Using Residuals," *Biometrika*, 62, 623–629.

BROWN, B. M., and KILDEA, D. G. (1979), "Outlier-Detection Tests and Robust Estimators Based on Signs of Residuals," *Communications in Statistics*, A8, 257–269.

CAMPBELL, N. A. (1978), "The Influence Function as an Aid in Outlier Detection in Discriminant Analysis," *Applied Statistics*, 27, 251–258.

——— (1980), "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," *Applied Statistics*, 29, 231–237.

CARROLL, R. J. (1980), "Robust Methods for Factorial Experiments With Outliers," *Applied Statistics*, 29, 246–251.

——— (1982), "Two Examples of Transformations When There Are Possible Outliers," *Applied Statistics*, 149–152.

CHAUVENET, W. (1960), *A Manual of Spherical and Practical Autonomy* (Vol. II, 5th ed.), New York: Dover.

CHHIKARA, R. S., and FEIVESON, A. L. (1980), "Extended Critical Values of Extreme Studentized Deviate Test Statistics for Detecting Multiple Outliers," *Communications in Statistics*, B9, 155–166.

COLLETT, D. (1980), "Outliers in Circular Data," *Applied Statistics*, 29, 50–57.

COLLETT, D., and LEWIS, T. (1976), "The Subjective Nature of Outlier Rejection Procedures," *Applied Statistics*, 25, 228–237.

CONOVER, W. J., BEMENT, T. R., and IMAN, R. L. (1979), "On a Method for Detecting Clusters of Possible Uranium Deposits," *Technometrics*, 21, 277–282.

COOK, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

——— (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

COOK, R. D., and BECKMAN, R. J. (1980), "Using *M*-Estimators to Identify Outliers," Los Alamos Report LA-UR 80-1418, Los Alamos, New Mexico.

COOK, R. D., HOLSCHUH, N., and WEISBERG, S. (1982), "A Note on an Alternative Outlier Model," *Journal of the Royal Statistical Society*, Ser. B, 44, 370–376.

COOK, R. D., and PRESCOTT, P. (1981), "On the Accuracy of Bonferroni Significance Levels for Detecting Outliers in Linear Models," *Technometrics*, 23, 59–63.

COOK, R. D., and WEISBERG, S. (1980), "Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495–508.

——— (1982), *Residuals and Influence in Regression*, New York: Chapman-Hall.

COX, D. R., and HINKLEY, D. V. (1974), *Theoretical Statistics*, London: Chapman-Hall.

COX, D. R., and SNELL, E. J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society*, Ser. B, 30, 248–275.

——— (1971), "On Test Statistics Calculated From Residuals," *Biometrika*, 58, 589–594.

DANIEL, C. (1960), "Locating Outliers in Factorial Experiments," *Technometrics*, 2, 149–156.

——— (1978), "Patterns in Residuals in the Two-Way Layout," *Technometrics*, 20, 385–395.

DANIELL, P. J. (1920), "Observations Weighted According to Order," *American Journal of Mathematics*, 42, 222–236.

DAVID, H. A. (1981), *Order Statistics* (2nd ed.), New York: John Wiley.

DAVIES, R. B., and HUTTON, B. (1975), "The Effects of Errors in the Independent Variables in Linear Regression," *Biometrika*, 62, 383–391.

DE ALBA, E., and VAN RYZIN, J. (1980a), "An Empirical Bayes Approach to Outliers," *Journal of Statistical Planning and Inference*, 4, 217–236.

——— (1980b), "A Empirical Bayes Approach to Outliers: Shifted Mean Case," in *Recent Developments in Statistical Inference and Data Analysis*, ed. K. Matusita, Amsterdam: North-Holland.

DE FINETTI, B. (1961), "The Bayesian Approach to the Rejection of Outliers," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley and Los Angeles: University of California Press, 199–210.

DEMPSTER, A. P., and GASKO-GREEN, M. (1981), "New Tools for Residual Analysis," *Annals of Statistics*, 9, 945–959.

DEMPSTER, A. P., and ROSNER, B. (1971), "Detection of Outliers," in *Statistical Decision Theory and Related Topics I*, ed. S. S. Gupta, New York: Academic Press, 161–180.

DIXON, W. J. (1950), "Analysis of Extreme Values," *Annals of Mathematical Statistics*, 21, 488–506.

——— (1951), "Ratios Involving Extreme Values," *Annals of Mathematical Statistics*, 22, 68–78.

——— (1953), "Processing Data for Outliers," *Biometrics*, 9, 74–89.

——— (1960), "Simplified Estimation From Censored Normal Samples," *Annals of Mathematical Statistics*, 31, 385–391.

——— (1962), "Rejection of Observations," in *Contributions to Order Statistics*, eds. A. E. Sarhan and B. G. Greenberg, New York: John Wiley, 299–321.

DIXON, W. J., and TUKEY, J. W. (1968), "Approximate Behavior of the Distribution of Winsorized *t* (Trimming/Winsorization 2)," *Technometrics*, 10, 83–98.

DOORNBOS, R. (1981), "Testing for a Single Outlier in a Linear Model," *Biometrics*, 37, 705–712.

DOORNBOS, R., and PRINS, H. J. (1958), "On Slippage Tests I. A General Type of Slippage Test and a Slippage Test for Normal Variates," *Indagationes Mathematicae*, 20, 38–55.

DRAPER, N. R., and HERTZBERG, A. M. (1979), "Designs to Guard Against Outliers in the Presence or Absence of Model Bias," *Canadian Journal of Statistics*, 7, 127–135.

DRAPER, N. R., and JOHN, J. A. (1980), "Testing for Three or Fewer Outliers in Two-Way Tables," *Technometrics*, 22, 9-15.

——— (1981), "Influential Observations and Outliers in Regression," *Technometrics*, 23, 21–26.

EDGEWORTH, F. Y. (1887), "On Discordant Observations," *Philosophical Magazine*, 23, Ser. 5, 364–375.

ELASHOFF, J. D. (1972), "A Model for Quadratic Outliers in Linear Regression," *Journal of the American Statistical Association*, 67, 478–485.

ELLENBERG, J. H. (1973), "The Joint Distribution of the Standardized Least Squares Residuals From a General Linear Regression," *Journal of the American Statistical Association*, 68, 941–943.

——— (1976), "Testing for a Single Outlier From a General Linear Model," *Biometrics*, 32, 637–645.

EPSTEIN, B. (1960a), "Tests for the Validity of the Assumption That the Underlying Distribution Life is Exponential: Part I," *Technometrics*, 2, 83–101.

——— (1960b), "Tests for the Validity of the Assumption That the Underlying Distribution of Life is Exponential: Part II," *Technometrics*, 2, 167–183.

FERGUSON, T. S. (1961), "On the Rejection of Outliers," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley and Los Angeles: University of California Press, 253–287.

FIELLER, N. R. J. (1976), "Some Problems Related to the Rejection of Outlying Observations," Unpublished Ph.D. thesis, University of Sheffield.

FISHER, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London*, A, 222, 309–368.

FOX, A. J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society*, Ser. B, 34, 340–363.

GALPIN, J. S., and HAWKINS, D. M. (1981), "Rejection of a Single Outlier in Two-or Three-Way Layouts," *Technometrics*, 23, 65–70.

GASTWIRTH, J. L. (1966), "On Robust Procedures," *Journal of American Statistical Association*, 61, 929–948.

GEBHARDT, F. (1964), "On the Risk of Some Strategies for Outlying Observations," *Annals of Mathematical Statistics*, 35, 1524–1536.

GENTLE, J. E. (1978), "Testing for Outliers in Linear Regression," in *Contributions to Survey Sampling and Applied Statistics, in Honor of H. O. Hartley*, ed. H. A. David, New York: Academic Press.

GENTLEMAN, J. F. (1980), "Finding the K Most Likely Outliers in Two-Way Tables," *Technometrics*, 22, 591–600.

GENTLEMAN, J. F., and WILK, M. B. (1975a), "Detecting Outliers in a Two-Way Table: I. Statistical Behavior of Residuals," *Technometrics*, 17, 1–14.

——— (1975b), "Detecting Outliers II. Supplementing the Direct Analysis of Residuals," *Biometrics*, 31, 387–410.

GLAISHER, J. W. L. (1873), "On the Rejection of Discordant Observations," *Monthly Notices of the Royal Astronomical Society*, 23, 391–402.

——— (1874), "Note on a Paper by Mr. Stone, 'On the Rejection of Discordant Observations'," *Monthly Notices of the Royal Astronomical Society*, 34, 251.

GNANADESIKAN, R., and KETTENRING, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection With Multiresponse Data," *Biometrics*, 28, 81–124.

GOLDSMITH, P. L., and BODDY, R. (1973), "Critical Analysis of Factorial Experiments and Orthogonal Fractions," *Applied Statistics*, 22, 141–160.

GOLDSTEIN, M. (1982), "Contamination Distributions," *The Annals of Statistics*, 10, 174–183.

GOULD, B. A. Jr. (1855), "On Peirce's Criterion for the Rejection of Doubtful Observations, With Tables for Facilitating Its Application," *Astronomical Journal*, 6, 81–83.

GREEN, R. F. (1974), "A Note on Outlier-Prone Families of Distributions," *Annals of Statistics*, 2, 1293–1295.

——— (1976), "Outlier-Prone and Outlier-Resistant Distributions," *Journal of the American Statistical Association*, 71, 502–505.

GRUBBS, F. E. (1950), "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, 21, 27–58.

——— (1969), "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, 11, 1–21.

GUTTMAN, I. (1973a), "Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity—A Bayesian Approach," *Technometrics*, 15, 723–738.

——— (1973b), "Premium and Protection of Several Procedures for Dealing With Outliers When Sample Sizes Are Moderate to Large," *Technometrics*, 15, 385–404.

GUTTMAN, I., and DUTTER, R. (1976), "Procedures for Investigating Outliers When Estimating in the General Univariate Linear Situation—Nonfull Rank Case," *Communications in Statistics*, A5, 819–835.

GUTTMAN, I., DUTTER, R., and FREEMAN, P. R. (1978), "Care and Handling of Univariate Outliers in the General Linear Model to Detect Spuriosity—A Bayesian Approach," *Technometrics*, 20, 187–193.

GUTTMAN, I., and KAHATRI, C. G. (1975), "A Bayesian Approach to Some Problems Involving the Detection of Spuriosity," in *Applied Statistics Symposium*, ed. R. P. Gupta, Amsterdam: North-Holland, 111–146.

GUTTMAN, I., and SMITH, D. E. (1969), "Investigation of Rules for Dealing With Outliers in Small Samples From the Normal Distribution I: Estimation of the Mean," *Technometrics*, 11, 527–550.

——— (1971), "Investigation of Rules for Dealing With Outliers on Small Samples From the Normal Distribution II: Estimation of the Variance," *Technometrics*, 13, 101–111.

HAMPEL, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

HARTER, H. L. (1978), *A Chronological Annotated Bibliography on Order Statistics* (Vol. 1: pre-1950), Washington, D.C.: U.S. Government Printing Office.

HARTLEY, H. O., and GENTLE, J. E. (1975), "Data Monitoring Criteria for Linear Models," in *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava, Amsterdam: North-Holland, 197–207.

HAWKINS, D. M. (1972), "Analysis of a Slippage Test for the Chi-Squared Distribution," *South African Statistical Journal*, 6, 11–17.

——— (1974), "The Detection of Errors in Multivariate Data Using Principal Components," *Journal of the American Statistical Association*, 69, 340–344.

——— (1978), "Letter to the Editor," *Technometrics*, 20, 218.

——— (1979), "Fractiles of an Extended Multiple Outlier Test," *Journal of Statistical Computation and Simulation*, 8, 227–236.

——— (1980), *Identification of Outliers*, London: Chapman and Hall.

HERZBERG, A. M., and ANDREWS, D. F. (1976), "Some Considerations in the Optimal Design of Experiments in Non-Optimal Situations," *Journal of the Royal Statistical Society Series*, Ser. B, 38, 284–289.

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

HODGES, J. L. Jr., and LEHMANN, E. L. (1963), "Estimates of Location Based on Rank Tests," *Annals of Mathematical Statistics*, 34, 598–611.

HOGG, R. V. (1967), "Some Observations on Robust Estimation," *Journal of the American Statistical Association*, 62, 1179–1186.

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

——— (1979), "Statistical Robustness: One View of Its Use in Applications Today," *The American Statistician*, 33, 108–115.

HOLLAND, P. W., and WELSCH, R. E. (1977), "Robust Regression Using Iteratively Reweighted Least-Squares," *Communications in Statistics*, A6, 813–827.

HUBER, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.

——— (1972), "The 1972 Wald Lecture Robust Statistics: A Review," *Annals of Mathematical Statistics*, 43, 1041–1067.

——— (1975), "Robustness and Designs," in *A Survey of Statistical Design and Linear Models*, ed. Jagdish N. Srivastava, Amsterdam: North-Holland.

——— (1977), *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics, Bristol, England: J. W. Arrowsmith Ltd.

——— (1981), *Robust Statistics*, New York: John Wiley.

——— (1983), "Minimax Aspects of Bounded-Influence Regression," *Journal of the American Statistical Association*, 78, 66–72.

IRWIN, J. O. (1925), "On a Criterion for the Rejection of Outlying Observations," *Biometrika*, 17, 238–250.

JAECKEL, L. A. (1971a), "Some Flexible Estimates of Location," *Annals of Mathematical Statistics*, 42, 1540–1552.

——— (1971b), "Robust Estimates of Location: Symmetry and Asymmetric Contamination," *Annals of Mathematical Statistics*, 42, 1020–1034.

JAIN, R. B. (1981a), "Percentage Points of Many-Outlier Detection Procedures," *Technometrics*, 23, 71–75.

——— (1981b), "Detecting Outliers: Power and Some Other Considerations," *Communications in Statistics. A. Theory and Methods*, 10, 2299–2314.

JAIN, R. B., and PINGEL, L. A. (1981a), "A Procedure for Estimating the Number of Outliers," *Communications in Statistics*, A10, 1029–1041.

——— (1981b), "On the Robustness of Recursive Outlier Detection Procedures to Non-normality," *Communications in Statistics*, A10, 1323–1334.

JEFFREYS, H. (1932), "An Alternative to the Rejection of Observations," *Proceedings of the Royal Society. London*, Ser. A, 137, 78–87.

JENNINGS, D. (1982), "Inference and Diagnostics for Logistic Regression," unpublished Ph.D. thesis, University of Minnesota, School of Statistics.

JOHN, J. A. (1978), "Outliers in Factorial Experiments," *Applied Statistics*, 27, 111–119.

JOHN, J. A., and DRAPER, N. R. (1978), "On Testing for Two Outliers or One Outlier in Two-Way Tables," *Technometrics*, 20, 69–78.

JOHN, J. A., and PRESCOTT, P. (1975), "Critical Values of a Test to Detect Outliers in Factorial Experiments," *Applied Statistics*, 24, 56–59.

JOHNSON, B. A., and HUNT, H. H. (1979), "Performance Characteristics for Certain Tests to Detect Outliers," *Proceedings of the Statistical Computing Section, American Statistical Association*, 247–249.

JOHNSON, W., and GEISSER, S. (1980a), "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis," University of Minnesota, School of Statistics Technical Report 365.

——— (1980b), "Assessing the Predictive Influence of Observations," Technical Report 355, University of Minnesota, School of Statistics.

JOSHI, P. C. (1972), "Some Slippage Tests of Mean for a Single Outlier in Linear Regression," *Biometrika*, 59, 109–120.

KABE, D. G. (1970), "Testing for Outliers From an Exponential Distribution," *Metrika*, 15, 15–18.

KALE, B. K. (1976), "Detection of Outliers," *Sankhyā*, Ser. B, 38, 356–363.

——— (1979), "Outliers—A Review," *Journal of the Indian Statistical Association*, 17, 51–67.

KALE, B. K., and SINHA, S. K. (1971), "Estimation of Expected Life in the Presence of an Outlier Observation," *Technometrics*, 13, 755–759.

KERR, R. A. (1982), "Test Fails to Confirm Cloud Seeding Effect," *Science*, 217, 234–236.

KIMBER, A. C. (1979), "Tests for a Single Outlier in a Gamma Sample With Unknown Shape and Scale Parameters," *Applied Statistics*, 28, 243–250.

KIMBER, A. C., and STEVENS, H. J. (1981), "The Null Distribution of a Test for Two Upper Outliers in an Exponential Sample," *Applied Statistics*, 30, 153–157.

KING, E. P. (1953), "On Some Procedures for the Rejection of Suspected Data," *Journal of the American Statistical Association*, 48, 531–533.

KITAGAWA, G. (1979), "On the Use of AIC for the Detection of Outliers," *Technometrics*, 21, 193–199.

KRASKER, W. S., and WELSCH, R. E. (1982), "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 595–604.

KRUSKAL, W. H. (1960), "Some Remarks on Wild Observations," *Technometrics*, 2, 1–3.

KUDO, A. (1956), "On Testing Outlying Observations," *Sankhyā*, 17, 67–76.

LAURENT, A. G. (1963), "Conditional Distribution of Order Statistics and Distribution of the Reduced *i*th Order Statistic of the Exponential Model," *Annals of Mathematical Statistics*, 34, 652–657.

LEWIS, T., and FIELLER, N. R. J. (1979), "A Recursive Algorithm for Null Distributions for Outliers: 1. Gamma Samples," *Technometrics*, 21, 371–376.

LIKÉS, J. (1966), "Distribution of Dixon's Statistic in the Case of an Exponential Population," *Metrika*, 11, 46–54.

LINGAPPAIAH, G. S. (1976), "Effect of Outliers on the Estimation of Parameters," *Metrika*, 23, 27–30.

LUND, R. E. (1975), "Tables for an Approximate Test for Outliers in Linear Regression," *Technometrics*, 17, 473–476.

MARKS, R. G., and RAO, P. V. (1978), "A Modified Tiao-Guttman Rule for Multiple Outliers," *Communications in Statistics*, A7, 113–126.

——— (1979), "An Estimation Procedure for Data Containing Outliers With a One-Directional Shift in the Mean," *Journal of the American Statistical Association*, 74, 614–620.

MARONNA, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scale," *The Annals of Statistics*, 4, 51–67.

McKAY, A. T. (1935), "The Distribution of the Difference Between the Extreme Observation and the Sample Mean on Samples of *n* From a Normal Universe," *Biometrika*, 27, 466–471.

McMILLAN, R. G. (1971), "Tests for One or Two Outliers in Normal Samples With Unknown Variance," *Technometrics*, 13, 87–100.

McMILLAN, R. G., and DAVID, H. A. (1971), "Tests for One of Two Outliers in Normal Samples With Known Variance," *Technometrics*, 13, 75–85.

MICKEY, M. R. (1974), "Detecting Outliers With Stepwise Regression," *Communications—UCLA Health Science Facility*, 1, 1–9.

MICKEY, M. R., DUNN, O. J., and CLARK, V. (1967), "Note on the Use of Stepwise Regression in Detecting Outliers," *Computers and Biomedical Research*, 1, 105–111.

MOUNT, K. S., and KALE, B. K. (1973), "On Selecting a Spurious Observation," *Canadian Mathematical Journal*, 16, 75–78.

MOSES, L. E. (1978), "Charts for Finding Upper Percentage Points of Student's *t* in the Range .01 to .00001," *Communications in Statistics. B. Simulation and Computation*, 7, 479–490.

MURPHY, R. B. (1951), "On Tests for Outlying Observations," unpublished Ph.D. thesis, Princeton University.

NACHTSHEIM, C. J. (1979), "Contributions to Optimal Design," unpublished Ph.D. thesis, University of Minnesota, School of Statistics.

NAIR, K. R. (1948), "The Distribution of the Extreme Deviate From the Sample Mean and Its Studentized Form," *Biometrika*, 35, 118–144.

NELDER, J. A., and WEDDERBURN, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370–384.

NEWCOMB, S. (1886), "A Generalized Theory of the Combination of Observations so as to Obtain the Best Result," *American Journal of Mathematics*, 8, 343–366.

NEYMAN, J., and SCOTT, E. L. (1971), "Outlier Proneness of Phenomena and of Related Distributions," in *Optimizing Methods in Statistics*, ed. J. S. Rustagi, New York: Academic Press.

OGRODNIKOFF, K. (1928), "On the Occurrence of Discordant Observations and a New Method of Treating Them," *Monthly Notices of the Royal Astronomical Society*, 88, 523–532.

O'HAGAN, A. (1979), "On Outlier Rejection Phenomena in Bayes Inference," *Journal of the Royal Statistical Society*, Ser. B, 41, 358–367.

PAULSON, E. (1952), "An Optimum Solution to the *K*-Sample

Slippage Problem for the Normal Distribution," *Annals of Mathematical Statistics*, 23, 610–616.

PEARSON, E. S., and CHANDRA SEKAR, C. (1936), "The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations," *Biometrika*, 28, 308–320.

PEIRCE, B. (1852), "Criterion for the Rejection of Doubtful Observations," *Astronomical Journal*, 2, 161–163.

——— (1878), "On Peirce's Criterion," *Proceedings of the American Academy of Arts and Sciences*, New Series, 5, 348–351.

PREGIBON, D. (1979), "Data Analytic Methods for Generalized Linear Models," unpublished Ph.D. thesis, University of Toronto, Dept. of Statistics.

——— (1981), "Logistic Regression Diagnostics," *Annals of Statistics*, 9, 705–724.

PRESCOTT, P. (1975), "An Approximate Test for Outliers in Linear Models," *Technometrics*, 17, 129–132.

——— (1977), "An Upper Bound for Any Linear Function of Normed Residuals," *Communications in Statistics*, B6, 83–88.

——— (1978), "Examination of the Behavior of Tests for Outliers When More Than One Outlier Is Present," *Applied Statistics*, 27, 10–25.

——— (1979), "Critical Values for a Sequential Test for Many Outliers," *Applied Statistics*, 28, 36–39.

——— (1980), "A Review of Some Robust Data Analyses and Multiple Outlier Detection Procedures," *Bulletin in Applied Statistics*, 141–158.

QUESENBERRY, C. P., and DAVID, H. A. (1961), "Some Tests for Outliers," *Biometrika*, 48, 379–390.

RIDER, P. R. (1933), "Criteria for Rejection of Observations," Washington University Studies—New Series, Science and Technology, No. 8.

ROHLF, F. J. (1975), "Generalization of the Gap Test for the Detection of Multivariate Outliers," *Biometrics*, 31, 93–101.

ROSNER, B. (1975), "On the Detection of Many Outliers," *Technometrics*, 17, 221–227.

——— (1977), "Percentage Points for the RST Many Outlier Procedure," *Technometrics*, 19, 307–312.

SARHAN, A. E., and GREENBERG, B. G. (1956), "Estimation of Location and Scale Parameters by Order Statistics From Singly and Doubly Censored Samples," *Annals of Mathematical Statistics*, 27, 427–451.

SAUNDERS, S. A. (1903), "Note on the Use of Peirce's Criterion for the Rejection of Doubtful Observations," *Monthly Notices of the Royal Astronomical Society*, 63, 432–436.

SCHWAGER, S. J., and MARGOLIN, B. (1982), "Detection of Multivariate Normal Outliers," *Annals of Statistics*, 10, 943–954.

SCHWEDER, T. (1976), "Some 'Optimal' Methods to Detect Structural Shifts or Outliers in Regression," *Journal of the American Statistical Association*, 71, 491–501.

SERFLING, R. S. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

SHAPIRO, S. S., and WILK, M. B. (1972), "An Analysis of Variance Test for the Exponential Distribution," *Technometrics*, 14, 355–370.

SINHA, S. K. (1972), "Reliability Estimation on Life Testing in the Presence of an Outlier Observation," *Operations Research*, 20, 888–894.

SIOTANI, M. (1959), "The Extreme Value of the Generalized Distances of the Individual Points in the Multivariate Normal Sample," *Annals of the Institute of Statistical Mathematics*, 10, 183–208.

SNEDECOR, G. W., and COCHRAN, W. G. (1967), *Statistical Methods* (6th ed.), Ames, Iowa: Iowa State University Press.

——— (1980), *Statistical Methods* (7th ed.), Ames, Iowa: Iowa State University Press.

SRIKANTAN, K. S. (1961), "Testing for a Single Outlier in a Regression Model," *Sankhyā*, Ser. A, 23, 251–260.

STEFANSKY, W. (1971), "Rejecting Outliers by Maximum Normal Residual," *The Annals of Mathematical Statistics*, 42, 35–45.

——— (1972), "Rejecting Outliers in Factorial Designs," *Technometrics*, 14, 469–478.

STIGLER, S. M. (1973), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920," *Journal of the American Statistical Association*, 68, 872–879.

——— (1975), "Studies in the History of Probability and Statistics. XXXIV. Napoleonic Statistics: The Work of Laplace," *Biometrika*, 62, 503–517.

STONE, E. J. (1868), "On the Rejection of Discordant Observations," *Monthly Notices of the Royal Astronomical Society*, 28, 165–168.

——— (1873), "On the Rejection of Discordant Observations," *Monthly Notices of the Royal Astronomical Society*, 34, 9–15.

STUDENT (1927), "Errors of Routine Analysis," *Biometrika*, 19, 151–164.

THOMPSON, W. R. (1935), "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation," *Biometrika*, 32, 214–219.

TIAO, G. C., and GUTTMAN, I. (1967), "Analysis of Outliers With Adjusted Residuals," *Technometrics*, 9, 541–559.

TIETJEN, G. L., and MOORE, R. H. (1972), "Some Grubbs-Type Statistics for the Detection of Several Outliers," *Technometrics*, 14, 583–597.

——— (1979), Correction to "Some Grubbs-Type Statistics for the Detection of Several Outliers" (Vol. 14, p. 583–597), *Technometrics*, 21, 396.

TIETJEN, G. L., MOORE, R. L., and BECKMAN, R. J. (1973), "Testing for a Single Outlier in Simple Linear Regression," *Technometrics*, 15, 717–721.

TIKU, M. L. (1975), "A New Statistic for Testing Suspected Outliers," *Communications in Statistics. A. Theory and Methods*, 4, 737–752.

——— (1977), Rejoinder to a comment on "A New Statistic for Testing Suspected Outliers," *Communications in Statistics*, A6, 1417–1422.

TIPPETT, L. H. C. (1925), "On the Extreme Individuals and the Range of Samples Taken From a Normal Population," *Biometrika*, 17, 364–387.

TUKEY, J. W. (1960), "A Survey of Sampling From Contaminated Distributions," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, eds. I. Olkin, S. G. Ghurey, W. Hoeffdine, W. G. Madow, and H. B. Mann, Standord: Stanford University Press, 448–485.

WALSH, J. E. (1950), "Some Nonparametric Tests of Whether the Largest Observations of a Set are Too Large or Too Small," *Annals of Mathematical Statistics*, 21, 583–592. (see also correction *Annals of Mathematical Statistics*, 24, 134.)

WEDDERBURN, R. W. M. (1976), "On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models," *Biometrika*, 63, 27–32.

WILKS, S. S. (1963), "Multivariate Statistical Outliers," *Sankhyā*, 25, 407–426.

WILLIAMS, D. A. (1973), "Letter to the Editors," *Applied Statistics*, 22, 407–408.

WINTLOCK, J. (1856), "On Professor Airy's Objections to Peirce's Criterion," *Astronomical Journal*, 4, 145–147.