The Annals of Statistics
1981, Vol. 9, No. 5, 945–959

NEW TOOLS FOR RESIDUAL ANALYSIS[1]

BY A. P. DEMPSTER AND M. GASKO-GREEN

Harvard University

Techniques are presented for stepwise selection of observations in order of discrepancy from a linear model, and for assigning a P value to each selected observation. A general class of criteria for discrepancy is defined, and leading members of the class are discussed. Two ways to define the sequence of P values for a general selection criterion are discussed, and their relative advantages are compared. The techniques are illustrated on several data sets using only the computationally simpler of the two approaches to P values.

1. Introduction. We discuss here ways to assess residuals from least squares fits to linear models. Specifically, we consider methods which remove observations one at a time from a least squares analysis, using a combination of residual size and influence on the fitted model to assess each particular observation as a candidate for removal. To define a procedure, one needs a rule for sequentially selecting observations for removal, and one needs a means of judging how far the sequential removal process should be carried. Our main goal is to provide new techniques for the latter process of judging nominated outliers, but first we review and study various selection rules.

Two related papers are Andrews and Pregibon (1978) and Cook (1977). Each advocates a specific selection rule, motivated by considerations of influence. Andrews and Pregibon discuss significance testing criteria while Cook's criterion is a direct measure of influence. We survey these and other selection rules in Section 3, where we introduce a general formulation of selection rules.

Data analysis should consider applying several different rules to a given data set. There can be no universal choice of a procedure because statistical analyses have a wide range of purposes calling for different and often context-dependent concerns about outliers. Also, as illustrated in Section 6, different rules may suggest different hypotheses about the structure of a single data set.

Statistical analysis may be directed to developing scientific knowledge, or to more technological or decision-oriented concerns. One scientific use of regression analysis is to search for relations, as part of a process variously described as exploratory data analysis or descriptive modelling. Part of description may be simply to identify observations which do not fit the relation defined by the bulk of the data. For this purpose it may be sufficient to select the observation with the largest absolute deviation, especially if deviations thus measured on a given scale have substantive meaning in relation to what is an important shift. This simplest criterion is represented by (3.4) below. Another concern is to identify and remove, or possibly downweight, outlying observations which have unusually large influence on a fitted model. Rules which incorporate such influence are illustrated by (3.5) and (3.7). The primary technological use of fitted regression models is to estimate or predict a value for the dependent variable Y associated with a new vector $\mathbf{X}$ of values of independent variables. In this case a good selection rule should reflect the influence of specific observations on estimated Y values associated with a range of $\mathbf{X}$ vectors. It is thus evident that many different selection rules can have a rational basis, depending on the circumstances.

Received June 9, 1980; revised January, 1981.

[1] This work was facilitated in part by National Science Foundation Grant MC77-27119.

AMS 1970 *subject classification.* Primary 62J05; secondary 62F07.

Key words. Linear models, regression, outliers, residuals, influence.

The novelty of our approach comes in the test criteria we suggest as aids for judging breaking points in the process of removing observations. These criteria are based on a concept introduced in Section 4 which appeals to P values calculated from distributions on pieces of an arc. We are not concerned with formal decision-analytic rules for selection and testing, but rather with heuristic and informal methods for revealing possible structure in data. We suggest P values that can be useful for this purpose, alongside more exploratory assessments of the effects of individual observations.

We consider only least squares regression analysis and associated inference techniques dependent on normality assumptions. Significance testing criteria could be developed for other distributional assumptions, and it should be anticipated that the resulting tests would give different answers. For example, if a long-tailed distribution of deviations from a regression plane is accepted as an accurate descriptive model, then extreme residuals picked up by a normal theory significance test are inappropriate as evidence of model failure. Thus model-dependence is an essential part of the problem. We do believe that methods based on normal null hypotheses have a useful place in applied statistics. In circumstances where normal plots of studentized residuals are performed in search of stragglers, our testing procedures are appropriate.

2. Notation and Background.

It will be useful to have facility with certain geometric quantities related to the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \tag{2.1}$$

where $\mathbf{Y}$ and $\mathbf{e}$ and $n \times 1$ vectors and $\mathbf{X}$ is an $n \times p$ array of rank p. The $1 + p + n$ columns of the matrix $(\mathbf{Y}, \mathbf{X}, \mathbf{I})$ define vectors in an n-dimensional Euclidean space $\mathscr{R}$ corresponding to the dependent variable, the p independent variables, and the n indicator variables for the observations, respectively.

Any vector $\mathbf{V}$ in $\mathscr{R}$ may be decomposed into $\mathbf{V}_X + \mathbf{V}_\perp$, where $\mathbf{V}_X$ is the component in the p-dimensional subspace $\mathscr{R}_X$ spanned by the column vectors of $\mathbf{X}$, while $\mathbf{V}_\perp$ is the component in the orthogonal $(n-p)$-dimensional subspace $\mathscr{R}_\perp$. It is well known that $\mathbf{X}\mathbf{X}^I$ and $\mathbf{I} - \mathbf{X}\mathbf{X}^I$ represent the complementary orthogonal projections $\mathbf{V} \rightarrow \mathbf{V}_X$ and $\mathbf{V} \rightarrow \mathbf{V}_\perp$, where $\mathbf{X}^I = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. In particular, we make use of

$$\mathbf{Y}_\perp = (\mathbf{I} - \mathbf{X}\mathbf{X}^I)\mathbf{Y} \tag{2.2}$$

and

$$\mathbf{I}_\perp = (\mathbf{I} - \mathbf{X}\mathbf{X}^I)\mathbf{I}, \tag{2.3}$$

where $\mathbf{Y}_\perp$ is the familiar residual vector after fitting $\mathbf{Y}$ to $\mathbf{X}$ by least squares, and the columns of $\mathbf{I}_\perp$ are similarly projections of the columns of $\mathbf{I}$ into $\mathscr{R}_\perp$. When we regard the columns of $\mathbf{I}$ as indicator variables for the n observations, we denote them by $\mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}$. The corresponding columns of $\mathbf{I}_\perp$ we denote by $\mathbf{1}_\perp, \mathbf{2}_\perp, \cdots, \mathbf{n}_\perp$.

From (2.3), we see that $\mathbf{I}_\perp$ can be interpreted both as a set of residual vectors and as the projection operator $\mathbf{I} - \mathbf{X}\mathbf{X}^I$ itself. And we see that $\mathbf{I}_\perp = \mathbf{I}_\perp^T$. Since $\mathbf{Y}_\perp$ and $\mathbf{I}_\perp$ are already in $\mathscr{R}_\perp$, it follows that projecting them into $\mathscr{R}_\perp$ leaves them unchanged, so that

$$\mathbf{Y}_\perp = \mathbf{I}_\perp^T\mathbf{Y}_\perp \tag{2.4}$$

and

$$\mathbf{I}_\perp = \mathbf{I}_\perp^T\mathbf{I}_\perp. \tag{2.5}$$

In Section 3 we compare different selection rules by expressing them in terms of $2n$ geometric quantities, namely, the squared lengths

$$g_i = \mathbf{i}_\perp^T\mathbf{i}_\perp \tag{2.6}$$

of the $\mathbf{i}_\perp$ for $1, 2, \cdots, n$, and the angles θ_i between $\mathbf{Y}_\perp$ and $\mathbf{i}_\perp$ expressible as

$$\cos \theta_\iota = q^{-1/2} g_\iota^{-1/2} \mathbf{i}_\perp^T \mathbf{Y}_\perp, \tag{2.7}$$

$i = 1, 2, \cdots, n$, where q is the squared length of $\mathbf{Y}_\perp$

$$q = \mathbf{Y}_\perp^T \mathbf{Y}_\perp. \tag{2.8}$$

The quantity q is the residual sum of squares. From (2.4) and (2.7) the residual vector is expressible as

$$\mathbf{Y}_\perp = q^{1/2}(g_1^{1/2}\cos \theta_1, g_2^{1/2}\cos \theta_2, \cdots, g_n^{1/2}\cos \theta_n)^T. \tag{2.9}$$

There are many close connections between the lengths and angles just defined and sampling distributions determined by the model (2.1) when the vector $\mathbf{e}$ consists of independent $N(0, \sigma^2)$ random quantities. Our significance tests use the fact that $\cos^2 \theta_\iota$ has a beta distribution with density proportional to $u^{-1/2}(i - u)^{1/2(n-p-3)}$, or equivalently that

$$t_i = (n - p - 1)^{1/2} \cot \theta_\iota \tag{2.10}$$

has the Student's-t distribution with $n - p - 1$ degrees of freedom. The quantity t_i is often called the ιth *studentized residual.*

The quantities g_i are determined solely by $\mathbf{X}$. They obviously satisfy $0 \le g_\iota \le 1$ and $\sum_{\iota=1}^n g_\iota = n - p$. From (2.5) and (2.6) we see that g_ι is the ιth diagonal element of $\mathbf{I}_\perp$, whereas, from (2.3), $1 - g_\iota$ is the ιth diagonal element of $\mathbf{X}\mathbf{X}^I$. Hoaglin and Welsch (1978) call $\mathbf{X}\mathbf{X}^I$ the "hat" matrix, because it carries $\mathbf{Y}$ into its least squares predicted value $\mathbf{Y}_X$, which they denote by $\hat{\mathbf{Y}}$. They argue that $1 - g_\iota$ is intepretable as a direct measure of the influence of the ιth observation on the least squares fit because $1 - g_\iota$ is the fraction of the ιth component of $\mathbf{Y}$ directly preserved in $\mathbf{Y}_X$.

Andrews and Pregibon (1978) stress interpretations based on the effects of removing the ιth observation from the regression. Suppose we denote by $q_{(\iota)}$ the residual sum of squares in the regression with the ιth observation removed, and by $\mathbf{X}_{(\iota)}$ the remainder of $\mathbf{X}$ after moving the ιth row. It can be shown that

$$\sin^2 \theta_i \frac{q_{(\iota)}}{q} \tag{2.11}$$

and

$$g_\iota = \frac{\det(\mathbf{X}_{(\iota)}^T \mathbf{X}_{(\iota)})}{\det(\mathbf{X}^T\mathbf{X})}. \tag{2.12}$$

As noted below, Andrews and Pregibon suggest the product $g_i \sin^2 \theta_\iota$ as the quantity governing selection of the most important outlier. Formula (2.12) has an important sampling distribution interpretation because $\det(\mathbf{X}^T\mathbf{X})^{-1}$ is proportional to the squared volume of the standard confidence ellipsoid for $\boldsymbol{\beta}$. Hence, a smaller g_i means a larger increase in the volume of the ellipsoid when the ιth observation is removed, and hence a larger influence of the ιth observation on the ,accuracy with which $\boldsymbol{\beta}$ is estimated.

Finally, it often helps the study of least squares operations to work with sweeping operations on the inner product matrix $(\mathbf{Y}, \mathbf{X}, \mathbf{I})^T(\mathbf{Y}, \mathbf{X}, \mathbf{I})$. As shown in Chapter 4 of Dempster (1969), the operator SWP applied to the ιth index of the 3rd part of the matrix has the effect of removing the ιth observation from the remaining parts corresponding to $\mathbf{Y}$ and $\mathbf{X}$. Also, the operator SWP applied to the part corresponding to $\mathbf{X}$, carries out least square analysis. Since these SWP operators commute, the first can be used to remove the ιth observation after the basic least squares calculations are performed. It is easy to check formulas (2.11) and (2.12) this way, but we omit details.

3. Selection Rules. Given data $\mathbf{Y}, \mathbf{X}$, a selection rule is a rule which nominates one of the n observations as most discrepant. Assuming that the rule is defined for general n, it can be applied successively to remove observations one at a time. That is, after removing the most discrepant observation, the same rule can be applied to the remaining sample of

$n - 1$ observations to obtain the second observation for removal, and so forth. In view of the repeated stepwise nature of the process, we need only discuss the first step.

We present rules covering a range of important purposes. We stress generally applicable rules, but conclude by illustrating the concept of a rule tailored to the needs of a particular hypothetical application. The comparsion of different selection rules is facilitated by use of the quantities θ_i and g_i defined in Section 2. Thus, to specify a rule we define a function $H(\cdot, \cdot)$ and nominate the mth observation as most discrepant if

$$H(\theta_m, g_m) > H(\theta_i, g_i) \ \forall \ i \neq m. \tag{3.1}$$

We assume that the function $H(\cdot, \cdot)$ is sufficiently well-behaved that the rule (3.1) determines m uniquely with probability one, and hence we ignore cases of ties.

Our examples satisfy the requirement that

$$H(\theta, g) = H(\pi - \theta, g) \ \forall \ (\theta, g), \tag{3.2}$$

which from (2.9) means that we judge positive and negative residuals of the same size symmetrically. Also, it appears reasonable to assume that $H(\theta, g)$ is monotone decreasing on $0 < \theta < \frac{\pi}{2}$ (increasing on $\frac{\pi}{2} < \theta < \pi$) for fixed g, on the heuristic grounds that two observations with the same g value should be compared by using the absolute size of their residuals. Our examples exhibit no firm rule, however, about the direction of monotonicity of $H(\cdot, \cdot)$ in g for fixed θ.

In terms of θ_i and g_i, the simplest rule ignores the g_i and uses

$$\begin{aligned} H(\theta, g) &= -\theta \qquad \text{on } 0 \le \theta \le \frac{\pi}{2} \\ &= -(\pi - \theta) \text{ on } \frac{\pi}{2} \le \theta \le \pi. \end{aligned} \tag{3.3}$$

Equivalently, $H(\theta, g)$ may be taken to be $\cos^2\theta$ or $\cot^2\theta$. From (2.10), the rule (3.3) is equivalent to choosing as most discrepant the observation with the largest absolute studentized residual $|t_i|$. From (2.9), another characterization of (3.3) is that components of the residual vector $\mathbf{Y}_\perp$ are scaled by factors proportional to $g_i^{-1/2}$ and then compared in absolute value.

The rule (3.3) is evidently motivated by consideration of sampling distributions. Each t_i has the same marginal distribution, namely, student's t on $(n - p - 1)$ degrees of freedom, under the null hypothesis that the components of $\mathbf{e}$ in (2.1) are independent $N(0, \sigma^2)$. Hence the use of this rule is most natural for a specific technical purpose: to test the null hypothesis of normal homoscedastic error terms against an alternative hypothesis which envisions tail contamination with large values. Such tests can be useful aids to statistical modelling, especially when followed by procedures which depend critically on normality in the far tail. As discussed in Section 1, however, we believe that other selection rules may be more directly relevant to the primary purposes of data analysis.

Another simple selection rule uses the largest absolute residual, which from (2.9) is equivalent to choosing

$$H(\theta, g) = g \cos^2 \theta. \tag{3.4}$$

As remarked in Section 1, rule (3.4) may be appealing in situations where the fitted model is accepted as reasonably accurate and the focus is on picking up for special study observations which do not appear to fit. Since the g_i tend to be close to 1 when n increases, criteria (3.3) and (3.4) tend to be similar, but (3.4) downweights influence.

Our third example is the rule proposed by Cook (1977) which can be shown to correspond to the choice

$$H(\theta, g) = \frac{1 - g}{g} \cos^2 \theta. \tag{3.5}$$

Cook derived (3.5) by calculating how far the least squares estimate of β moves when the ith observation is removed, where distance is determined by radii of the standard confidence ellipsoid for β based on all of the data. We suggest an alternative derivation keyed to the use of regression for prediction.

Suppose that $\mathbf{b}$ and $\mathbf{b}_{(i)}$ denote the least squares estimates of β based on all n observations, and on $n-1$ observations after removal of the ith observation, respectively. Suppose that it is intended to apply the estimated regression coefficients to predict $\mathbf{Y}$ for a set of k $\mathbf{X}$'s represented by a $k \times p$ matrix $\mathbf{X}^*$. Then a plausible measure of the influence of the ith observation in the prediction is

$$\{\mathbf{X}^*\mathbf{b} - \mathbf{X}^*\mathbf{b}_{(i)}\}^T\{\mathbf{X}^*\mathbf{b} - \mathbf{X}^*\mathbf{b}_{(i)}\}. \tag{3.6}$$

If $\mathbf{X}^*$ is chosen to be $\mathbf{X}$, then selection based on (3.6) can be shown to be equivalent to (3.5).

Because the factor $(1-g)/g$ in (3.5) increases as influence increases, it is clear that (3.5) moves away from (3.3) in the direction of rewarding influence, whereas (3.4) moves oppositely. Both (3.4) and (3.5) possess the property that an observation with $\theta_i = 0$ or π need not be selected as most outlying, since maximizing $\cos^2\theta_i$ need not outweigh the other factor involving g_i. This property deserves attention because, when $\cos^2\theta_i = 1$, removal of the ith observation results in a perfect fit to the remaining $n-1$ observations, i.e. $q_{(i)} = 0$, as shown by (2.11). It may be more disturbing in the case of (3.4) than (3.5) because the selected observation is both less influential and $\mathbf{Y}_\perp$ has larger angle with $\pm i_\perp$. Replacing $\cos^2\theta$ by $\cot^2\theta$ in (3.5) removes this disturbing property and the resulting rule maximizes the standardized sum of squared changes in all regression coefficients when an observation is omitted. This diagnostic is suggested in Welsch and Peters (1978) and discussed in Belsley, Kuh and Welsch (1980).

The fourth general rule uses the Andrews and Pregibon rule

$$H(\theta, g) = -g\sin^2\theta. \tag{3.7}$$

From (2.11) and (2.12) it follows that (3.7) is equivalent to

$$H(\theta_i, g_i) = \det\{(\mathbf{Y}_{(i)}, \mathbf{X}_{(i)})^T(\mathbf{Y}_{(i)}, \mathbf{X}_{(i)})\}/\det\{(\mathbf{Y}, \mathbf{X})^T(\mathbf{Y}, \mathbf{X})\}, \tag{3.8}$$

which is mathematically appealing since the right side of (3.8) is the generalized variance proposed by Wilks (1932). Also, both influence and small θ are rewarded. Note that here an observation with $\theta_i = 0$ is automatically selected. A weakness of (3.7) is that the Wilks generalized variance has no direct connection with the primary goals of data analysis. Also, we note in Section 4 a mathematical awkwardness in relation to our proposed significance tests.

The rules (3.3), (3.4), (3.5), and (3.7) provide a kit of general purpose tools. We conclude by illustrating a possible special purpose tool. If the primary purpose is to look at a particular regression coefficient, say the first component β_1 of β, than a rule based on (3.6) would be appropriate, when $k = 1$ and $X^* = (1, 0, 0, \cdots, 0)$. Such a rule is not determined by θ_i and g_i alone. Cook (1979) extends (3.5) to situations in which a number of linearly independent combinations of the elements of β are of interest. The corresponding rule is also not determined by θ_i and g_i alone.

4. Testing the First Selected Observation.

Our main purpose in this paper is to propose the use of a sequence of P values corresponding to a sequence of observations chosen by stepwise application of a selection rule. We will describe in Section 5 two ways to obtain such a sequence of P values for any given selection rule. The two methods adopt the same initial P value, but differ in the second and later values of the sequence. In Section 4 we discuss the common initial P value, but not the later P values.

The null hypothesis defining the first P value is that $\mathbf{Y}$ is random according to the model (2.1), where $\mathbf{X}$ is fixed, and $\mathbf{e}$ is a vector of n independent $N(0, \sigma^2)$ random variables. The test statistic is tied to the selection rule, and hence depends on the values of $H(\theta_i, g_i)$

for $i = 1, 2, \cdots, n$. Since $\mathbf{X}$ is regarded as fixed, the g_i are likewise fixed, so the test statistic depends only on the random θ_i, i.e. is determined by the direction of $\mathbf{Y}_\perp$ in $\mathscr{R}_\perp$. Under the null model, the direction of $\mathbf{Y}_\perp$ is distributed independent of $q = \mathbf{Y}_\perp^T\mathbf{Y}_\perp$, and hence the null distribution can equivalently be described either in the unconditional way introduced above, or as a conditional distribution given q. Taking the latter view, we regard the null model as a uniform distribution over the surface $\mathscr{S}$ of the sphere of radius $q^{1/2}$ centered at the origin in $\mathscr{R}_\perp$.

Our proposed P value is defined by a further conditioning which reduces the sample to a subset of a certain one-dimensional arc on $\mathscr{S}$. Before describing our highly conditional test, we describe in contrast the conventional approach to testing a nominated largest outlier, which uses the uniform distribution on $\mathscr{S}$. The idea is to reject the null hypothesis if $H(\theta_m, g_m)$ defined in (3.1) is too large. To this end, a P value is the probability under the null hypothesis that a random $H(\theta_m, g_m)$ exceeds the observed value. This probability is the volume of a region on $\mathscr{S}$ which typically is the union of n pairs of polar caps around the axes $\pm i_\perp$, $1 \leq i \leq n$. The ith pair consists of all points where $H(\theta_i, g_i)$ exceeds the observed $H(\theta_m, g_m)$, $i = 1, 2, \cdots, n$ respectively. Usually the $2n$ caps are not disjoint and their intersection cannot be easily computed, hence, computing the volume of their union is not trivial. Andrews (1971) and Andrews and Pregibon (1978) suggest an approximation to this volume.

Turning now to our more conditional approach, we define several steps of conditioning beyond reducing the sample space to $\mathscr{S}$. For demonstration we use Figure 4.1 where $R_\perp$ is the three demensional Euclidean space, the selection rule is (3.3), and only $\mathbf{i}_\perp$, $\mathbf{j}_\perp$, and $\mathbf{m}_\perp$, three of the unit residual vectors, are considered. First, we condition on the observed m, the index of the first selected observation. This reduces the sample space to a pair of opposite subregions of $\mathscr{S}$ which we denote by Ω_{+m} and Ω_{-m}, where the sign ς_m of the mth component of $\mathbf{Y}_\perp$ is $+1$ on Ω_{+m} and -1 on Ω_{-m}. Note that θ_m is closer to 0 or to π according as $\mathbf{Y}_\perp \in \Omega_{+m}$ or $\mathbf{Y}_\perp \in \Omega_{-m}$, respectively. Second, we condition on the observed ς_m, reducing the sample space to Ω_{+m} or Ω_{-m} according as the observed ς_m is $+1$ or -1. Finally, we condition on the signs and absolute sizes relative to each other of the $n - 1$ residuals

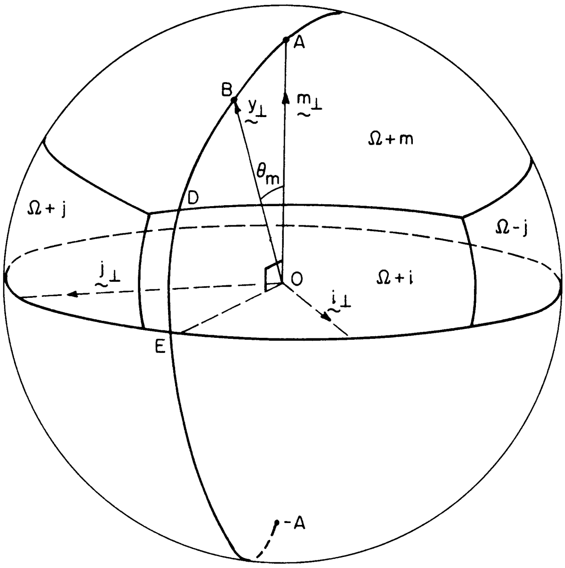
A
B
$\underset{\sim}{y}_\perp$
$\underset{\sim}{m}_\perp$
Ω+m
θ_m
D
Ω+j
Ω-j
$\underset{\sim}{j}_\perp$
O
Ω+i
$\underset{\sim}{i}_\perp$
E
-A

FIG. 4.1. *A geometrical representation of selection rule* (3.3) *and the conditional sample space.*

obtained from the reduced least squares analysis omitting the mth observation. The final condition by itself, being equivalent to the direction of the component of $\mathbf{Y}_\perp$ in the subspace of $\mathscr{R}_\perp$ orthogonal to $\mathbf{m}_\perp$, ($\overrightarrow{OE}$ in Figure 4.1), restricts the sample space to a semi-circle $\mathscr{C}$ on $\mathscr{S}$. $\mathscr{C}$ can be described as the half great circle joining two opposite poles A and $-A$ of $\mathscr{S}$ and passing through, B, the observed $\mathbf{Y}_\perp$, where the two opposite poles are defined by the intersection of $\mathscr{S}$ with the line formed by extending $\mathbf{m}_\perp$ in both directions. Hence, under the full conditioning, the sample space for the test is

$$\mathscr{C}_{\jmath_m m} = \mathscr{C} \cap \Omega_{\jmath_m m}. \tag{4.1}$$

In the example of Figure 4.1., $\mathscr{C}_{\jmath_m m}$ is the arc from A to D where at D the semicircle $\mathscr{C}$ intersects the boundary of $\Omega_{\jmath_m m}$.

The argument for conditioning on m is that some observation must be nominated by the selection rule even when none is out of line, so that the value of m does not in itself convey any information bearing on the significance or lack thereof of the first selected observation. The argument for conditioning on $\jmath_m$ is likewise that the sign of the mth residual carries no information about the significance of the residual. The final step of conditioning is introduced largely to obtain a simply understood and easily computable criterion. The information excluded by this conditioning is used both in the selection and testing procedures in subsequent steps of the proposed sequential process, as described in Section 5.

The null distribution along the arc $\mathscr{C}$ is characterized by Student's-t distribution (2.10), and hence the conditional null distribution over the restricted arc $\mathscr{C}_{\jmath_m m}$ is characterized by a restriction of the t-distribution to $\mathscr{C}_{\jmath_m m}$. Significance is judged by calculating the conditional probability that a random $\cos^2\theta_m$ would exceed the observed $\cos^2\theta_m$.

In most applications, the intersection (4.1) defines a single connected subarc of $\mathscr{C}$ running from the pole $\theta_m = 0$ or $\theta_m = \pi$ to the boundary of $\Omega_{\jmath_m m}$. In such cases, the P value is simply the ratio of two tail areas calculated from the t distribution (2.10), namely the measure of the arc from the pole to the observed θ_m divided by the measure of the full subarc $\mathscr{C}_{\jmath_m m}$. The main computational task is to find the value of θ_m at the boundary of $\Omega_{\jmath_m m}$. For certain selection rules, $\mathscr{C}_{\jmath_m m}$ may be less well-behaved, depending on the particular data set. For example, as noted in Section 3, a rule such as the Cook rule which chooses $H(\theta, g)$ to be a weighted $\cos^2\theta$ need not select the ith observation even when $\cos^2\theta_i = 1$. Even more paradoxically, the rule may select this ith observation for certain $\cos^2\theta_i < 1$, so that $\mathscr{C}_{\jmath_i}$ becomes a subarc which excludes the pole. Other kinds of pathological behavior can arise where $\mathscr{C}_{\jmath_m m}$ consists of a union of disjoint segments of $\mathscr{C}$. Such pathologies will be discussed in a later report. Here we simply note that the principle for computing P values defined in the preceding paragraph still applies, but the calculations become more difficult because the end points of various subarcs of $\mathscr{C}$ must be located.

We conclude our discussion of the first P value with a cautionary note. A small P value does indicate that an improbable event has occurred, where the class of improbable outcomes making up the event are those with larger $H(\theta_m, g_m)$ than the observed $H(\theta_m, g_m)$. We point out, however, that an unsurprising conditional P values, say .2 or .6, does not rule out the possibility that $H(\theta_m, g_m)$ is significantly large according to the unconditional test mentioned at the beginning of Section 4. The reason is that the conditional P value, being the specified ratio of tail areas assesses the magnitude of the most extreme observed discrepancy not on an absolute scale, but rather its excess over the next most extreme discrepancy in the data. While in this ratio the numerator may be very small, so may the denominator be small.

For example, suppose there are two extremely discrepant observations in the data with distinctly larger values on a chosen criterion $H(\theta, g)$ than the rest of the data. Initially the observation with the largest value on the criterion is nominated most discrepant. Although the corresponding unconditional P value may be very small the conditional P value can be of a moderate, unsurprising size. This is because the presence of a second outlier implies that the denominator of the conditional P value, being the probability measure associated with $\mathscr{C}_{,_m m}$ only slightly exceeds the probability measure association with $\mathscr{C}_{,_m m}$ having a larger value of $H(\theta_m, g_m)$ than observed, which is the numerator. From the definition of the complete sequence of conditional P values, in Section 5, it follows that for the above example the second element in the sequence will be small, suggesting that the first two selected observations are an outlying subset.

Thus, the first conditional P value should be interpreted alone only when it is small as suggesting that the selected most discrepant observation is significantly more discrepant than the next most discrepant observation. A moderate to large conditional P value would not be interpreted in isolation from the next P values in the sequence. The sequences of P values discussed in Section 5 are intended to provide markers between subsets of observations removed in sequence.

5. Two Sequences of *P* Values. Each of the two sequences to be described has advantages and disadvantages. The first sequence is computationally relatively easy, while the second sequence requires multiple integrations which are feasible only with a Monte Carlo simulation for each P value. Apart from computational ease, there is an important conceptual difference between the two sequences. Neither sequence is fully satisfactory, for reasons which we shall explain, and both can usefully be reported. For computational reasons, the examples in Section 6 report only the first sequence. A later report will illustrate the second sequence, and will analyze conditions under which the two approaches give similar or different results.

The first sequence is most easily described. Upon selecting observations one at a time, a P value is computed at each step in the sequential process as described in Section 4. That is, after removing the first observation, we pretend that the reduced sample of $n - 1$ observations is like an original sample. We select a second observation, the most discrepant in the reduced sample, and compute a second P value, as described in Sections 3 and 4, respectively, and so on. Each P value relates to a null hypothesis that the remaining data at that step are generated from the correspondingly reduced model (2.1). Analogous conditioning to that described in Section 4 implies that the kth P value can be computed using the t distribution on $n - p - k$ degrees of freedom. Each conditional P value is uniformly distributed under the respective null hypothesis which relates to the correspondingly reduced model.

The P values in the first sequence are not strictly independent nor marginally uniform under the full null hypothesis that all of the components of $\mathbf{e}$ in the model (2.1) are independent $N(0, \sigma^2)$. Accordingly, we have defined, and studied, a second set $P_2^{(1)}$, $P_2^{(2)}$, $\cdots$, $P_2{}^{(n-p)}$ which are independent uniform random variables, conditional on both the order of selection of the observations and the signs of the residuals of each selected observation at the stage when it is selected. Since a considerable amount of notation is required to give a precise definition of the $P_2^{(k)}$, we postpone details to a later paper which compares the two sequences using numerical studies. However, the basic idea is to define $P_2^{(k)}$ from the precise conditional sampling distributions under the full null hypothesis using the same angles as in the first sequence. Unlike in the first sequence where the cotangents of the angles are simply assigned t distributions, in computing the second sequence marginals of joint sampling densities need to be derived. The latter densities do not have a simple analytic representation but can be straightforwardly simulated.

Which of the sequences $P_1^{(1)}$, $P_1^{(2)}$, $\cdots$, or $P_2^{(1)}$, $P_2^{(2)}$, $\cdots$ is more appropriate in practice? The use envisaged for either sequence is to mark off groups of observations which appear to fit the model (2.1) notably less well than the remaining observations. For example, if the sequence were to begin .62, .48, .001, .8, .01, $\cdots$ the suggestion would be that removing the first three observations implies a better fitting model with the remaining $n - 3$, and that the next two form another group that worsens the fit of the remaining $n - 5$. The method is intended to serve as a suggestive diagnostic.

As long as the full null hypothesis remains tenable, it would seem to be preferable to use $P_2^{(r)}$, since these are independently uniformly distributed. But after some observations are rejected, like the first three in the above hypothetical example, then the original null hypothesis becomes irrelevant. At this point, $P_2^{(4)}$ is computed from a sampling null distribution that does not relate well to the then current null hypothesis, namely, the hypothesis that observations $m^{(r)}$ for $r = 4, 5, \cdots$ fit the model (2.1) given that $m^{(1)}$, $m^{(2)}$, $m^{(3)}$ have been rejected. Indeed, if $m^{(1)}$, $m^{(2)}$, $m^{(3)}$ are judged to be extreme outliers, then the corresponding ordinary t distribution on $n - p - 4$ degrees of freedom becomes approximately relevant, i.e., $P_1^{(4)}$ is more appropriate than $P_2^{(4)}$.

Thus, we conclude that both sequences have merit and neither should be preferred to the other.

6. Examples We illustrate our proposed methodology by applying the four selection rules (3.3)-(3.5), (3.7) to analyze three data sets two of which are discussed in Andrews and Pregibon (1978). For each data set we have tabulated the first 15 observations ordered by discrepancy, along with the corresponding sequence of P values for each of the four selection rules. Our results indicate that rules (3.3) and (3.4) based on studentized residuals and absolute residual sizes, respectively, suggest similar structure of data except when a moderate to large residual is associated with an excessively influential observation on the fitted model. The other two rules, (3.5) and (3.7) based on criteria suggested by Cook (1977) and Andrews and Pregibon (1978), respectively, also suggest similar structure in the data but different from that of the first two rules. As emphasized in Section 2 we do not recommend one type of rule over the others because all may provide valuable insight into the data. In Section 7 we present the computational methods for selection, testing, and removal of selected observations.

Example 1. Mickey, Dunn, and Clark (1967). The data describe observations on 21 children where the response is the Gessel Adaptive score and the independent variable is age in months at first word. Figure 6.1 shows a plot of the data, and Table 6.1 shows the ordered observations and the corresponding P values using each of the four selection rules.

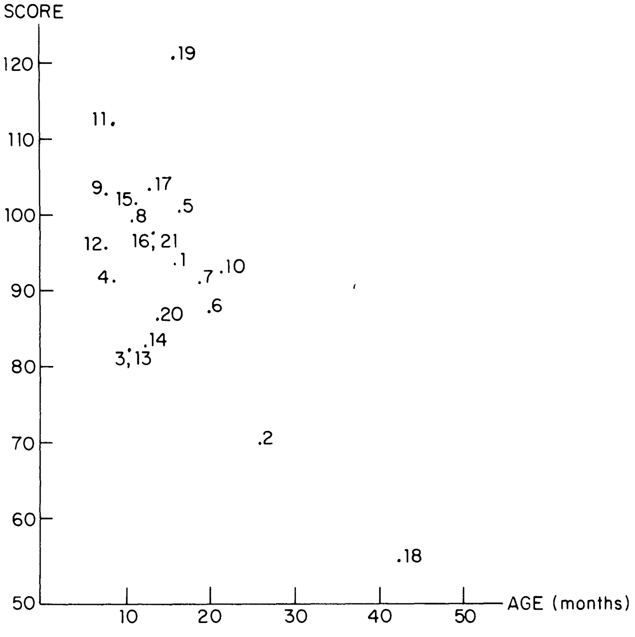

FIG. 6.1. *Children's score in an aptitude test vs. age.*

TABLE 6.1
Children's scores
Observations listed in a descending order of discrepancy with corresponding P values using four selection rules

Studentized Residuals (3.3)		**Absolute Residual Size (3.4)**		**Andrews and Pregibon (3.7)**		**Cook's Criterion (3.5)**	
observation	P value	observation	P value	observation	P value	observation	P value
19	.0225	19	.0221	18	.4091	18	.6375
3	1.0000	3	1.0000	2	.2165	2	.2592
13	.5670	13	.5733	19	.1270	19	.1503
14	.5766	14	.5766	11	.7070	11	.2077
20	.5036	20	.4886	6	.4416	6	.8185
4	.2530	4	.2242	10	.8288	9	.7510
2	.8677	2	.9777	7	.6166	3	1.0000
12	.2514	12	.2611	5	.5291	13	.5621
5	.8927	5	.8188	1	.9376	14	.5702
11	.7334	11	.7888	9	.1908	20	.5620
10	.9960	10	.9932	17	.2855	4	.1448
17	.0359	17	.0364	12	.9222	12	.3353
7	.5112	7	.4951	4	.9467	17	.7691
15	.1434	15	.1410	15	1.0000	5	.3093
1	.7363	1	.6961	8	.5590	10	.3970

Figure 6.1 suggests that observations 18, 2, and 19 in this order stand out. The rules (3.3) and (3.4) imply an identical order with two similar sequences of P values which identify two significant gaps in the data. The first gap comes between the initially selected observation 19 and the remainder of the data, in agreement with the analysis of Mickey et al. based on reduction in residual sums of squares which is an identical selection criterion to (3.4). The second gap is at the 12th step of the process when the remaining $21 - 12 - 9$ observations fit very closely to a straight line which is similar to the least squares fit by all the data.

The rules (3.5) and (3.7) do not find significant gaps in the data, suggesting that the observations are compatible among themselves and with the linear model. Both rules select in their first three steps the above mentioned outstanding observations, i.e., 18, 2, and 19 in that order, but the large corresponding P values suggest that 18 and 2 merely extend the domain of the model. Combining the results of all three rules suggests that observation 19 selected for its large studentized and absolute residual size has no significant influence on the fitted model conforming to what Andrews and Pregibon call an outlier that does not matter.

TABLE 6.2
Data from oxidizing ammonia plant

No.	x_1	x_2	x_3	y	No.	x_1	x_2	x_3	y	No.	x_1	x_2	x_3	y
1.	80	27	89	42	**8.**	62	24	93	20	**15.**	50	18	89	8
2.	80	27	88	37	**9.**	58	23	87	15	**16.**	50	18	86	7
3.	75	25	90	37	**10.**	58	18	80	14	**17.**	50	19	72	8
4.	62	24	87	28	**11.**	58	18	89	14	**18.**	50	19	79	8
5.	62	22	87	18	**12.**	58	17	88	13	**19.**	50	20	80	9
6.	62	23	87	18	**13.**	58	18	82	11	**20.**	56	20	82	15
7.	62	24	93	19	**14.**	58	19	93	12	**21.**	70	20	91	15

x_1 = *air flow*; x_2 = *cooling water inlet temperature*; x_3 = *acid concentration*; y = *stack loss*

TABLE 6.3
Oxidizing ammonia data
Observations listed in a descending order of discrepancy with corresponding P values using four selection rules

Studentized Residuals (3.3)		**Absolute Residual Size (3.4)**		**Andrews and Pregibon (3.7)**		**Cook's Criterion (3.5)**	
observation	P value	observation	P value	observation	P value	observation	P value
21	.1596	21	.2561	21	.0613	21	.0291
4	.0529	4	.0468	4	.2992	4	.9833
3	.5157	3	.5352	2	.8412	2	.9724
1	.0673	1	.1567	1	.5346	3	.8695
13	.2246	13	.2862	3	.0004	1	.0023
20	.9806	20	.7474	13	.4675	13	.3219
2	.9768	14	.6066	14	.9353	14	.3927
14	.1967	6	.6550	20	.2539	8	.9415
8	.3762	15	.9266	8	.3732	20	.3816
16	.4734	5	.9407	16	.6470	16	.4361
12	.8801	9	.5655	15	.9926	12	.6389
19	.5500	7	.5040	12	.5827	7	.9189
7	.0251	19	.3470	11	.4316	19	.0131
5	.2415	11	.1634	17	.7234	15	.4008
15	.5833	18	.2694	19	.0157	6	.3279

Example 2. Daniel and Wood (1975), Brownlee (1965). The data of this example describe the operation of a plant oxidizing ammonia to nitric acid. The rows in Table 6.2 refer to observations taken on 21 successive days of operation while the columns refer to three input variables and a response variable which (inversely) measures the efficiency of the system. Table 6.3 presents the results of our analysis.

Daniel and Wood (1971, Chapter 5) investigate the data using residual plots from least squares fit to various subsets of the observations, realizing the masking effect by which days 1–4 conspire to hide a significantly single most discrepant observation. Following careful examination of the data, they find observations 21, 4, 3, and 1 to be outliers. They remark that the system underwent a transient stage during the initial days of operation and that it makes sense to include observation 2 among the outliers.

Turning to Table 6.3 we see that the selection rules (3.5) and (3.7) identify initially observation 21 as significantly most discrepant and a second last significant gap after selecting the observations taken on the initial four days, including the second day. The results of the two rules (3.3) and (3.4) are similar to each other and to the findings of Daniel and Wood. In this case examination of the raw data in Table 6.3 does not easily suggest the results.

Example 3. Doll (1955). The data include observations on yearly cigarette consumption per capita in 1930 and deaths per million in 1950 from 11 countries. Figure 6.2 shows a plot of the data and the least squares fitted lines using all observations and excluding observation 7 denoted by solid and broken lines, respectively. Observation 7 stands out and has the largest residual.

Table 6.4 summarizes the analyses using the four selection rules. All rules select initially observation 7 and its P value is significant. Indeed, omitting observation 7 the fitted line is steeper. The second and third selected observations by all rules are 10 and 11 respectively. Both (3.3) and (3.4), not weighting influence, do not suggest these as an outlying subset. However, both (3.7) and (3.5) considering influence find the two observations jointly and sequentially, respectively, significant outliers.

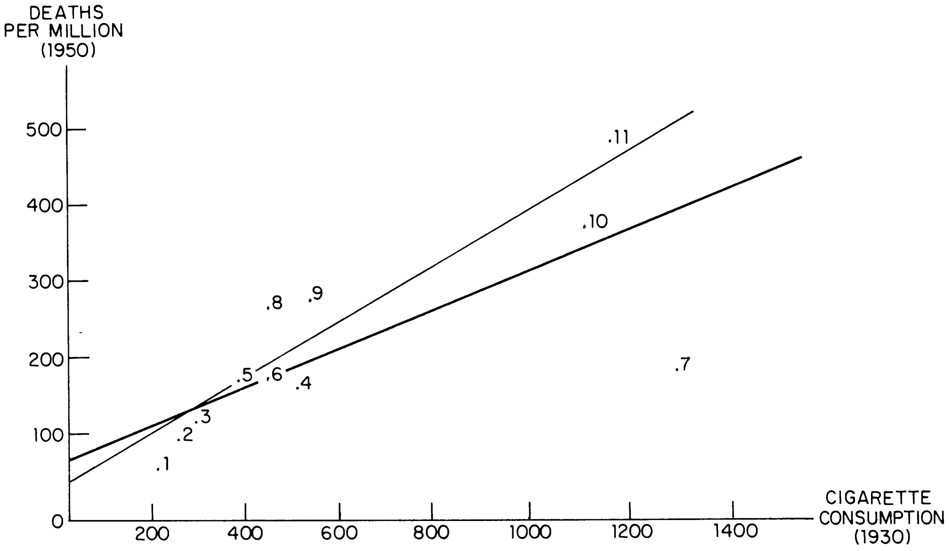

FIG. 6.2. *Smoking data.*

TABLE 6.4
Smoking Data
Observations listed in a descending order of discrepancy with corresponding P-values using four selection rules

Studentized Residuals (3.3)		**Absolute Residual Size (3.4)**		**Andrews and Pregibon (3.7)**		**Cook's Criterion (3.5)**	
observation	P value	observation	P value	observation	P value	observation	P value
7	.0027	7	.0022	7	.0156	7	.0201
10	.2293	10	.3183	10	.3338	10	.0641
11	.8873	8	.7726	11	.0817	11	.0996
4	.1120	4	.6724	4	.2178	8	.7883
6	.6316	9	.2442	6	.9540	9	.6272
8	.0940	5	.2814	8	.1177	4	.3826
2	.3042	1	.0268	9	.8301	6	.0619
5	.5785	6	.0322	2	.4383	1	.7925

The fourth and successive moderate to large P values in all sequences suggest that among the remaining 8 observations there are no more outliers. The few small P values at the end of the sequence can be ignored because they relate to situations when a totality of only 4, 3, or 2 observations are considered.

7. Computational Methods. We used a straightforward computational strategy to carry out the computations involved in selecting a most discrepant observation, assessing the corresponding P value and eliminating the selected observation from the data in preparation of a reduced sample for the next step of the process. This computational strategy is not the most economical in storage space; however, we end this section describing two alternative computational strategies calling for decreasing amounts of storage locations compensated by increasing amounts of calculations. Description of software can be found in an unpublished Ph.D. dissertation, "*P* Values for Sequentially Selected Outliers," by M. Gasko-Green (Statistics Department, Harvard University, Cambridge, MA).

The simplest strategy assumes that at every step of the process, say k, the numbers in corresponding rows to the reduced data in the array $(\mathbf{Y}_\perp^{(k)}, \mathbf{I}_\perp^{(k)})$ are directly accessible. The $k+1$ step array, $(\mathbf{Y}_\perp^{(k+1)}, \mathbf{I}_\perp^{(k+1)})$ is computed by sweeping $(\mathbf{Y}_\perp^{(k)}, \mathbf{I}_\perp^{(k)})$ along the $m^{(k)}$ column of $\mathbf{I}^{(k)}$ which corresponds to the observation being removed, as described in Chapter 4 of Dempster (1969).

In view of the stepwise repeated nature of our process it suffices to describe the computational aspect of the intial step of the process while in successive steps the computations are identical using the corresponding quantities to the reduced samples.

We show that the computations involved both in selecting and testing the first most discrepant observation use some elements of the inner product matrix $(\mathbf{Y}_\perp, \mathbf{I}_\perp)^T(\mathbf{Y}_\perp, \mathbf{I}_\perp)$. The identities (2.4) and (2.5) imply that except for $q = \mathbf{Y}_\perp^T\mathbf{Y}_\perp$, the (1.1) element of this matrix, these elements need not be calculated once we have computed and can directly access the elements of $(\mathbf{Y}_\perp, \mathbf{I}_\perp)$.

In order to compute any general selection rule of the form described in Section 3 we need the quantities g_ι and either $\cos^2\theta_i$ or $\sin^2\theta_i$, $i = 1, 2, \cdots, n$. From (2.9) it follows that g_ι is the ith diagonal element of $\mathbf{I}_\perp$. From (2.5) and (2.7) $\cos^2\theta_i$ is the ratio of the element in ith row of $\mathbf{Y}_\perp$ squared to the product $q \times g_i$, and $\sin^2\theta_\iota = 1 - \cos^2\theta_\iota$.

The main computational task involved in computing the first conditional P value is to compute the range Φ of angles θ_m associated with the conditional sample space $\mathscr{C}_{\mathfrak{s}_m m}$. The semicircle $\mathscr{C}$ defined in Section 4 is expressible as

$$\text{(7.1)} \qquad \mathbf{Y}_\perp(\phi) = q^{-1/2}(\mathbf{v}_m \cos\phi + \mathbf{u}_m \sin\phi), \qquad 0 \le \phi \le \pi$$

where $\mathbf{v}_m$ is a unit vector in the direction of $\mathbf{m}_\perp$, $\mathbf{v}_m = g_m^{-1/2}\mathbf{m}_\perp$ and $\mathbf{u}_m$ is a unit vector in the direction of the component of $\mathbf{Y}_\perp(\phi)$ in the subspace of $\mathscr{R}_\perp$ orthogonal to $\mathbf{m}_\perp$. The angles θ_i defined in Section 2, corresponding to $\mathbf{Y}_\perp(\phi)$ are denoted by $\theta_i(\mathbf{Y}_\perp(\phi))$, $i = 1, 2, \cdots, n$, respectively. Recalling (4.1), $\mathscr{C}_{\mathfrak{s}_m m}$ is the intersection of the $n - 1$ arc subsets of $\mathscr{C}$, all contained in $\Omega_{\mathfrak{s}_m m}$ and each consists of $\mathbf{Y}_\perp(\phi)$ such that m is selected over specific j, $j \neq m$. Hence Φ can be expressed as

$$\text{(7.2)} \qquad \Phi = \bigcap_{\substack{j=1\\ j\neq m}}^{n} \{\phi : H(\theta_m(\mathbf{Y}_\perp(\phi), g_m) \\ > H(\theta_j(\mathbf{Y}_\perp(\phi), g_j), \mathbf{Y}_\perp(\phi) \text{ of the form (7.1)}, \mathfrak{s}_m \cos\phi > 0 \text{ and } 0 < \phi < \pi\}$$

First we show how to compute (7.2) for rules of the form $H(\theta_i, g_i) = w_i \cos^2\theta_i$ and later for rule $H(\theta_i, g_i) = -w_i \sin^2\theta_i$, where $w_i = w(g_i) > 0$ is a known function.

$H(\theta, g) = \cos^2\theta$. Expressing $\mathbf{Y}_\perp(\phi)$ as in (7.1) shows $\cos^2\theta_m(\mathbf{Y}_\perp(\phi)) = \cos^2\phi$ and

$$\text{(7.3)} \qquad \left.\begin{aligned} \cos^2\phi_j(\mathbf{Y}_\perp(\phi)) = \frac{(\mathbf{Y}_\perp(\phi)^T\mathbf{j}_\perp)^2}{q \cdot q_j} = (\mathbf{v}_m^T\mathbf{v}_j \cos\phi + \mathbf{u}_m^T\mathbf{v}_j \sin\phi)^2 \\ j \neq m \end{aligned}\right\}$$

Using (7.3), given j, $1 \le j \le n$, $j \neq m$, the corresponding range, Φ_j, defined in (7.2) consists of $0 \le \phi \le \pi$, $\mathfrak{s}_m \cos\phi > 0$ such that

$$\text{(7.4)} \qquad w_m \cos^2\phi > w_j(\mathbf{v}_m^T\mathbf{v}_j \cos\phi + \mathbf{u}_m^T\mathbf{v}_j \sin\phi)^2$$

In order to simplify notation denote $a = \mathbf{v}_m^T\mathbf{v}_j$ and $b = \mathbf{u}_m^T\mathbf{v}_j$. Simple arithmetic manipulations imply that (7.4) is equivalent to

$$\text{(7.5)} \qquad f(\cot\phi) = [w_m - w_j a^2]\cot^2\phi - 2w_j ab \cot\phi - w_j b^2 > 0.$$

Condition (7.5) together with $\cot\phi \gtrless 0$ as $\mathfrak{s}_m \gtrless 0$ and the equality $\cot\alpha = \cot(\pi - \alpha)$ imply that the jth region Φ_j is expressible as

$$\Phi_j = \begin{cases} (\phi^{**}, \phi^{*}) & \mathfrak{s}_m = +1 \\ (\pi - \phi^{*}, \pi - \phi^{**}) & \mathfrak{s}_m = -1, \end{cases} \tag{7.6}$$

where

$$\phi^* = \text{arc cot} \frac{b}{\pm\sqrt{\dfrac{w_m}{w_j}} - \mathfrak{s}_m a} \qquad \text{with} \pm \text{as} \quad b \gtrless 0$$

and

$$\phi^{**} = \begin{cases} 0 & & \text{if } w_m > w_j a^2 \\ \text{arc cot} \dfrac{b}{\mp\sqrt{\dfrac{w_m}{w_j}} - \mathfrak{s}_m a} & \text{with} \mp \text{as } b \gtrless 0, & \text{otherwise} \end{cases}$$

Computing ϕ^*, ϕ^{**} is simple in view of the earlier remark about the inner product matrix, specifically

$$a = \frac{(\mathbf{I}_\perp)_{m,j}}{\sqrt{(\mathbf{I}_\perp)_{m,m} \cdot (\mathbf{I}_\perp)_{j,j}}}$$

and

$$b = \frac{(\mathbf{Y}_\perp)_j - \dfrac{(\mathbf{Y}_\perp)_m}{(\mathbf{I}_\perp)_{m,m}} \cdot (\mathbf{I}_\perp)_{m,j}}{\sqrt{(\mathbf{I}_\perp)_{j,j}}\sqrt{q - \dfrac{(\mathbf{Y}_\perp)_m^2}{(\mathbf{I}_\perp)_{m,m}}}}$$

where $(\mathbf{Y}_\perp)_m$ is the mth component of $\mathbf{Y}_\perp$, and $(\mathbf{I}_\perp)_{i,j}$ is the (i, j) element of $\mathbf{I}_\perp$.

$H(\theta, g) = -w \sin^2\theta$. Using the equality $\sin^2\alpha = 1 - \cos^2\alpha$ and (7.3), algebraic manipulations imply that the jth region in (7.2) consist of angles ϕ such that $\cot \phi \gtrless 0$, as $\mathfrak{s}_m \gtrless 0$ and

$$f(\cot \phi) = (1 - a^2)\cot^2\phi - 2ab \cot \phi + 1 - b^2 - \frac{w_m}{w_j} > 0. \tag{7.7}$$

Condition (7.7) together with $\cot \phi \gtrless 0$ as $\mathfrak{s}_m \gtrless 0$ imply that Φ_j is expressible

$$\Phi_j = \begin{cases} (0, \phi^*) \cup \left(\phi^{**}, \dfrac{\pi}{2}\right) & \text{if} \quad \mathfrak{s}_m = +1 \\ \left(\dfrac{\pi}{2}, \pi - \theta^{**}\right) \cup (\pi - \theta^*, \pi) & \text{if} \quad \mathfrak{s}_m = -1 \end{cases}$$

where, writing $\Delta = b^2 - (1 - a^2)\{1 - (w_m/w_j)\}$,

$$\phi^* = \begin{cases} \pi/2 & \text{if} \quad \Delta < 0 \\ \text{arc cot}\left(\dfrac{\mathfrak{s}_m ab + \sqrt{\Delta}}{1 - a^2}\right) & \text{otherwise,} \end{cases}$$

and

$$\phi^{**} = \begin{cases} \text{arc cot}\left(\dfrac{\mathfrak{s}_m ab - \sqrt{\Delta}}{1 - a^2}\right) \text{ if } (1 - a^2)\left(1 - \dfrac{w_m}{w_j}\right) < b^2 < 1 - \dfrac{w_m}{w_j} \\ \dfrac{\pi}{2} \qquad \text{otherwise} \end{cases}$$

The second computational strategy assumes that, $(\mathbf{X}^{IT}, \mathbf{Y}_{\perp})$, a smaller matrix is directly accessible, where $\mathbf{X}^{I}$ is defined in Chapter 2. When needed we compute the required elements $\mathbf{I}_{\perp}$, using $\mathbf{I}_{\perp} = \mathbf{I} - \mathbf{X}^{IT}\mathbf{X}^{T}$ which increases the amount of computations involved. Initially $(\mathbf{X}^{IT},\mathbf{Y}_{\perp})$ is obtained by sweeping the first P columns from (X,Y). The working array corresponding to the reduced sample after eliminating observation m from the data is obtained by sweeping on $\mathbf{m}_{\perp}$ from $(\mathbf{X}^{IT}, \mathbf{m}_{\perp}, \mathbf{Y}_{\perp})$ where $m_{\perp}$ need to be calculated.

The third computational strategy, the most economical in storage space, assumes that at each step we have the arrays $(\mathbf{X}^{T}\mathbf{X})^{-1}$, $\hat{\beta}$ and the residual sum of squares q which may be stored, or at least conceived, as the single matrix

$$\mathbf{Q} = \begin{bmatrix} (\mathbf{X}^{T}\mathbf{X})^{-1} & \hat{\beta} \\ \beta^{T} & q \end{bmatrix}.$$

At the initial step, $\mathbf{Q}$ is obtained by sweeping the first p columns from the inner product matrix $(\mathbf{XY})^{T}(\mathbf{XY})$. The elements of $\mathbf{I}_{\perp}$ and $\mathbf{Y}_{\perp}$ are computed as needed by first computing $\mathbf{X}^{I}$ and thereafter proceeding with the second algorithm. The next step Q matrix corresponding to the reduced data without observation m is obtained by sweeping the last column from

$$\begin{bmatrix} Q & \hat{\beta}_{\cdot m} \\ \hat{\beta}^{T}_{\cdot m} & g_m \end{bmatrix}$$

where $\hat{\beta}_{\cdot m} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{i}^{T}$ and $\mathbf{X}_{i}^{T}$ is the ith row of $\mathbf{X}$ and g_m is the (m, m) element of $\mathbf{I}_{\perp}$.

REFERENCES

ANDREWS, D. F. (1971). Significance tests based on residuals. *Biometrica* **58** 139–148.

ANDREWS, D. F. and PREGIBON, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc. Ser. B* **40** 85–94.

BELSLEY, A. B., KUH, E. AND WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

BROWNLEE, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering.* Wiley, New York.

COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19** 15–18.

COOK, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.* **74** 169–174.

DANIEL, D. and WOOD, F. S. (1971). *Fitting Equations to Data.* Wiley, New York.

DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Wesley, Reading, MA.

DOLL, R. (1955). Etiology of lung cancer. *Advances in Cancer Research* **3**.

HOAGLIN, D. C. and WELSCH, R. E. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.* **32** 17-22.

MICKEY, M. R., and DUNN, O. J., and CLARK, V. (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomed. Res.* **1** 105–111.

WEISBERG, S. (1980). *Applied Linear Regression.* Wiley, New York.

WELSCH, R. E. AND PETERS, C. S. (1978). Finding Influential Subsets of Data in Regression Models. Proc. of Comp. Sci. and Statistics: 11th Annual Symposium on the Interface. Institute of Statistics, North Carolina State Univ.

WILKS, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24** 471–494.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
ONE OXFORD STREET
CAMBRIDGE, MASSACHUSETTS