Cross-validatory Choice and Assessment of Statistical Predictions

By M. Stone

University College London

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 5th, 1973, Professor J. Gani in the Chair]

Summary

A generalized form of the cross-validation criterion is applied to the choice and assessment of prediction using the data-analytic concept of a prescription. The examples used to illustrate the application are drawn from the problem areas of univariate estimation, linear regression and analysis of variance.

Keywords: CROSSVALIDATION; PRESCRIPTION; DOUBLECROSS; CHOICE OF VARIABLES; MODELMIX; PREDICTION; UNIVARIATE ESTIMATION; MULTIPLE REGRESSION; ANALYSIS OF VARIANCE

1. Introduction

This paper will be concerned with applications of a *cross-validation* criterion to the choice and assessment of statistical prediction. The concept of such assessment† is an old one. In its most primitive but nevertheless useful form, it consists in the controlled or uncontrolled division of the data sample into two subsamples, the choice of a statistical predictor, including any necessary estimation, on one subsample and then the assessment of its performance by measuring its predictions against the other subsample. An example of controlled division is provided by the cautious statistician who sets aside a randomly selected part of his sample without looking at it and then plays without inhibition on what is left, confident in the knowledge that the set-aside data will deliver an unbiased judgment on the efficacy of his analysis. Examples of *uncontrollable* division are to be found in the criminological studies of Simon (1971) where a "construction sample" and a "validation sample" may be separated by an interval of several years. Larson (1931) employed random division of the sample in an educational multiple-regression study to investigate the "shrinkage of the coefficient of multiple correlation" between the fabrication and trying out of predictors. [Amusingly, Larson's statistically reckless inferences from the size of the observed shrinkages led Wherry (1931) to formulate the by now well-known "adjusted multiple correlation coefficient" $\bar{R}^2$.] Horst (1941), in his fascinating study of the prediction of success in marriage, found a "drop in predictability" between an "original" sample and a "check" sample that depended strongly on the method of construction of the predictor. Mosier, Cureton, Katzell and Wherry (1951) contributed separate papers to a "Symposium: The need and means of cross-validation" which added further emphases and suggested some solutions. Nicholson (1960) addressed the shrinkage phenomenon theoretically and proposed an alternative measure of prediction efficiency. Herzberg (1969) made a detailed theoretical and numerical study of predictor construction methods, using cross-validatory assessment.

† The term *assessment* is preferred to *validation* which has a ring of excessive confidence about it.

In their authorship study, Mosteller and Wallace (1963, 1964) used assessment measures on a sequence of subsamples in order to *select* words of high discrimination ability. Anderson *et al.* (1972) presented some dramatic examples of the fall-off in performance of predictors between "index" sample and "follow-up" sample, examples that included the use of sophisticated procedures such as forward and stepwise selection.

The refinement of this type of assessment that gives us our cross-validation criterion appears to have been developed by Lachenbruch following a suggestion in Mosteller and Wallace (1963). Key references are Hills (1966), Kerridge in discussion of the same, Lachenbruch and Mickey (1968) and Cochran (1968). The context of all this work is that of discrimination, as is the application described by Mosteller and Tukey (1968) to whom, however, we are indebted for the first clear general statement of the refinement. Their description of what they term "simple cross-validation" is worth reproducing:

> "Suppose that we set aside one individual case, optimize for what is left, then test on the set-aside case. Repeating this for every case squeezes the data almost dry. If we have to go through the full optimization calculation every time, the extra computation may be hard to face. Occasionally we can easily calculate either exactly or to an adequate approximation what the effect of dropping a specific and very small part of the data will be on the optimized result. This adjusted optimized result can then be compared with the values for the omitted individual. That is, we make one optimization for all the data, followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual, followed by one test of that individual. When practical, this approach is attractive."

The assessment or cross-validation criterion that is implied by this quotation is therefore one that corresponds to division of the sample (size n) into a "construction" subsample (size $n-1$) and a "validation" subsample (size 1) in all (n) possible ways. The key to our present approach is just a small step from the stance of the quotation but a crucial one nevertheless. We bring in the question of choice of predictor and employ the implied cross-validation criterion in a way that *integrates the procedures of choice and assessment*. (The precise technique of this integration will be explained in Section 2.)

An illustration of the choice component of this integrated method was outlined in the discussion of Efron and Morris (1973) and is discussed in a slightly variant form in Example 3.2 below. Independently, Geisser (1974) has arrived at the same method for choice of estimator in the very same context. Geisser once described the approach as of "predictive jack-knife type". The confusion between cross-validation and jack-knifing is easily understood since both employ the device of omission of items one or more at a time. Gray and Schucany (1972, pp. 125–136) appear to initiate the confusion in their description of Mosteller and Tukey's sophisticated, simultaneous juggling act with the two concepts. The component of jack-knifing that sharply distinguishes it from cross-validation is its manufacture of pseudovalues for the reduction of bias. This is not to suggest that the jack-knife (bias reducer) may not be found eventually to have some application to cross-validation.

Section 2 develops the general framework and definitions that are then illustrated in a sequence of examples in Section 3. Example 3.1 deals with the well-worn problem of univariate estimation from a single sample without concomitant information; we here cover the unsophisticated use of the sample mean as well as the more realistic exploitation of the order statistic. Examples 3.2 and 3.3 show that cross-validatory choice can easily generate estimators similar in behaviour to the most fashionable Bayesian and decision theoretic ones. With data from 27 Earth satellites as a restraining influence, Example 3.4 considers a cross-validatory approach to the choice of variables in "linear regression", using the "double-cross" technique. An alternative method of application is the "model-mix" technique of Example 3.5. The important special case of symmetry, which applies to standard balanced experimental designs, is examine in Example 3.6.

While it is the purpose of this paper to be mainly illustrative and expository, some theoretical overtones and undertones are struck in Section 4.

2. General Framework

Our data set is, we suppose, a sample S of measurements (x, y) on each of n items, where x and y may be quite general. Thus we may write

$$S = \{(x_i, y_i) \mid i = 1, \ldots, n\}.$$

Consider a new item for which only the x-value has been measured and it is required to predict the y-value by $\hat{y}$, a function of x and S. We will suppose that our starting point is a *prescription* (class of predictors)

$$\{\hat{y}(x;\, \alpha, S) \mid \alpha \in \mathscr{A}\}, \tag{2.1}$$

where the dependence of $\hat{y}(x;\, \alpha, S)$ on x and S is prescribed; moreover, the dependence on S is specified for samples of size n, $n-1$ and $n-2$, while $\mathscr{A}$ is independent of n. The element of choice in (2.1) lies in allowing S to determine α. Choice of α as a function of S will be described as "estimation of α" although it should not be supposed that the concept of a "true" value of α has any meaning.

Example 2.1. If x and y are real numbers, we may have the prescription

$$\hat{y}(x;\, \alpha, S) = \bar{y} + \alpha\left\{\frac{\Sigma(x_i - \bar{x})\, y_i}{\Sigma(x_i - \bar{x})^2}\right\}(x - \bar{x}) \tag{2.2}$$

with $\mathscr{A} = [0, 1]$.

It is not the purpose of this paper to prejudge the form of such prescriptions; the case for adopting a prescriptive approach must rest on the appeal and usefulness of examples such as those that follow the present general statement.

However, it is reasonable to enquire how one arrives at a prescription in any particular problem. A tentative answer is that, like a doctor with his patient, the statistician with his client must write his prescription only after careful consideration of the reasonable choices, suggested *a priori* by the nature of the problem or even by current statistical treatment of the problem type. Just as the doctor should be prepared for side-effects, so the statistician should monitor and check the execution of the prescription for any unexpected complications. A prescription should be accorded a status quite distinct from that of a model in the customary approach, although there is an informal connection that may be exploitable in the choice of prescription (see Example 3.4 below). A prescription is neither true nor false; it is better to say that, in a broad sense, it either succeeds or fails.

We informally develop the *method of cross-validatory choice of* α and the *method of cross-validatory assessment of this choice* as follows:

I. A *naive choice* of α is the value $\alpha^0(S) \in \mathscr{A}$ that minimizes

$$\bar{L}(\alpha) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ \alpha, S)]$$

where $L[y, \hat{y}]$ is some selected loss function of $\hat{y}$ as a predictor of the actual value y.

Example 2.2. In Example 2.1 with $L[y,\hat{y}] = (y-\hat{y})^2$, we find $\alpha^0(S) = 1$, corresponding to the customary least-squares fitting procedure.

II. *Naive assessment of this naive choice* would employ $\bar{L}(\alpha^0(S))$, that is, the average over the n items in S of $L[y, \hat{y}(x;\ \alpha^0(S), S)]$.

Example 2.3. For Example 2.1 and quadratic L, we find that $\bar{L}(\alpha^0(S))$ equals RSS/n where RSS denotes the customary residual sum of squares.

Even if $\alpha^0(S)$ is not naive in a pejorative sense, such description of this assessment is, in general, justified by the existence of the shrinkage phenomenon already mentioned.

III. *Cross-validatory assessment of the naive choice* would employ

$$C^0 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ \alpha^0(S_{\backslash i}), S_{\backslash i})]$$

where $S_{\backslash i}$ denotes the sample S with the ith item omitted and $\alpha^0(S_{\backslash i})$ is the naive choice of α based on $S_{\backslash i}$, that is, the value of α minimizing

$$\bar{L}_{\backslash i}(\alpha) \overset{\text{def}}{=} \frac{1}{(n-1)}\sum_{\backslash i} L[y_j, \hat{y}(x_j;\ \alpha, S_{\backslash i})].$$

where $\sum_{\backslash i}$ denotes summation omitting the ith item. C^0 is strictly relevant to the prediction problem for samples of size $n-1$ rather than n but this conservatism may be a small price to pay for the increase in realism over $\bar{L}(\alpha^0(S))$ that might be expected as a result of the shrinkage phenomenon.

Example 2.4. For Example 2.1 and L quadratic, a straightforward calculation shows that

$$C^0 = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i-\hat{y}_i)^2}{[1-n^{-1}-(x_i-\bar{x})^2/\Sigma(x_i-\bar{x})^2]^2},$$

where $\hat{y}_i$ is the least-square fitted value of y_i. Comparison of C^0 with

$$\bar{L}(\alpha^0(S)) = \frac{1}{n}\Sigma(y_i-\hat{y}_i)^2$$

shows that cross-validation gives greater weight to the residuals for large $|x_i-\bar{x}|$.

IV. *Cross-validatory choice of* α is the value $\alpha^\dagger(S) \in \mathscr{A}$ that minimizes

$$C(\alpha) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ \alpha, S_{\backslash i})]. \tag{2.3}$$

Example 2.5. For Example 2.1 and L quadratic, $\alpha^\dagger(S)$ is rather complicated but for the special case of n even and an equal number, $\frac{1}{2}n$, of x_i's at each of two values we obtain

$$\alpha^\dagger(S) = \left(1 - \frac{(n-1)(1-r^2)}{(n-1)(n-3)r^2+1}\right)^+,$$

where r is the sample correlation coefficient of x and y and $(z)^+ = \max(z, 0)$. This example will be generalized in Section 3 but we may here note its effect when deployed in (2.2). When r^2 is close to 1, (2.2) will be close to the least-squares prediction but, when r^2 approaches zero, there will be a drastic "flattening" of the least-squares regression coefficient to the neighbourhood of zero.

V. *Cross-validatory assessment of this cross-validatory choice* employs

$$C^\dagger \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ \alpha^\dagger(S_{\backslash i}), S_{\backslash i})], \tag{2.4}$$

where $\alpha^\dagger(S_{\backslash i})$ is the cross-validatory choice of α based on $S_{\backslash i}$, that is, the value of α minimizing

$$C_{\backslash i}(\alpha) \overset{\text{def}}{=} \frac{1}{(n-1)}\sum_{\backslash i} L[y_j, \hat{y}(x_j;\ \alpha, S_{\backslash ij})],$$

where $S_{\backslash ij}$ denotes the sample S with the ith and jth items omitted. Thus we see that cross-validatory assessment of a cross-validatory choice involves a "two-deep" analysis (cf. Mosteller and Tukey, 1968, p. 147).

Example 2.6. For Example 2.1 and L quadratic, the expression for $C^\dagger$ is hopelessly complex and the assessment must be computer-based.

For convenience of reference, Fig. 1 gives the shape of the cross-validatory paradigm.

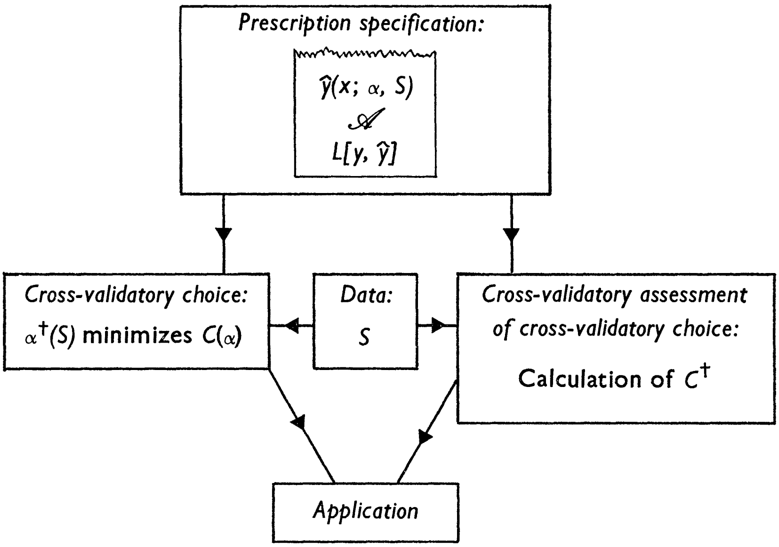

FIG. 1. The cross-validatory paradigm.

The following remark reveals that the dependence of $\hat{y}$ on its third argument, S, is at the root of cross-validatory choice.

Remark. If $\hat{y}(x;\ \alpha, S)$ *is independent of* S *then* $\alpha^\dagger(S) \equiv \alpha^0(S)$.

However, even when the condition of the remark holds, it may still prove useful to calculate $C^\dagger$. In general, the principal motivation for dismissing naive choice, $\alpha^0(S)$, in favour of cross-validatory choice, $\alpha^\dagger(S)$, is the belief, which can at present only be supported by numerical examples, that in most problems $C^\dagger$ will be appreciably less than C^0. In other words, cross-validatory choice will be a better statistical method.

There are two ramifications of the general theory that may be conveniently described at this stage. Both will find later application.

VI. *Two-stage cross-validatory choice of* α; *"double-cross"*. It may, in some cases, be convenient to treat α as having two components; $\alpha = (a, b)$ where $a \in A$ and $b \in B(a)$ (for example, see Example 3.4 below). With a fixed, $b^\dagger(S_{\backslash i}, a)$ denotes the value of b that minimizes

$$\frac{1}{(n-1)} \sum_{\backslash i} L[y_j, \hat{y}(x_j;\ (a,b), S_{\backslash ij})].$$

Then $a^{\dagger\dagger}(S)$ denotes the value of a that minimizes

$$C^\dagger(a) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ (a, b^\dagger(S_{\backslash i}, a)), S_{\backslash i})]. \tag{2.5}$$

Finally choose $b = b^{\dagger\dagger}(S)$ in $B(a^{\dagger\dagger}(S))$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ (a^{\dagger\dagger}(S), b), S_{\backslash i})].$$

Write $\alpha^{\dagger\dagger}(S) = (a^{\dagger\dagger}(S), b^{\dagger\dagger}(S))$ for what may be called the two-stage cross-validatory choice of α or "double-cross". The cross-validatory assessment of $\alpha^{\dagger\dagger}(S)$ is the same as for $\alpha^\dagger(S)$, namely, by calculation of

$$C^{\dagger\dagger} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} L[y_i, \hat{y}(x_i;\ \alpha^{\dagger\dagger}(S_{\backslash i}), S_{\backslash i})], \tag{2.6}$$

where $\alpha^{\dagger\dagger}(S_{\backslash i})$ is obtained from $S_{\backslash i}$ just as $\alpha^{\dagger\dagger}(S)$ is from S. The value of double-cross will be exemplified in a case (Example 3.4) where $\hat{y}(x;\ \alpha, S)$ is independent of S. In such a case, simple cross-validatory choice would yield the naive estimator; double-cross is designed to avoid this outcome.

VII. There is no very good reason why the loss function used in the definition of $\alpha^\dagger(S)$ or $\alpha^{\dagger\dagger}(S)$ should be the same as that used in its assessment. (The latter loss function should reasonably be considered fixed.) Indeed we have some evidence (see Example 3.1, Table 1) that this flexibility might be exploited. It is not unreasonable to suggest that the choice of L might itself be made by a variant of VI, that is, by attaching a to L rather than to $\hat{y}$. However, we have not explored this possibility.

3. EXAMPLES

The following examples have been chosen to illustrate the range of application of the above general method. The prescriptions, that is, forms of (2.1), have been chosen, in one or two cases at least, with a deliberate carelessness, to see whether the cross-validatory method has an unpleasant and unacceptable face.

Example 3.1. *The location of a single univariate sample.* In this problem, x is absent and y is real. So we are required to choose α in $\hat{y}(\alpha, S)$ where $S = \{y_1, \ldots, y_n\}$. In this case, we may regard $\hat{y}(\alpha^\dagger(S), S)$ as a cross-validatory estimator of the location of the y values. Four prescriptions are examined:

(a) With $\hat{y}(\alpha, S) \equiv \alpha$, $\mathscr{A} = R^1$ and L quadratic, we have $\alpha^\dagger(S) = \bar{y}$ and $C^\dagger = [n/(n-1)]s^2$ where $s^2 = \Sigma(y_i - \bar{y})^2/(n-1)$. In this case $\alpha^0(S) = \bar{y}$ also.

(b) With $\hat{y}(\alpha, S) = \alpha\bar{y}$ (a prescription that gives $y = 0$ a special status), $\mathscr{A} = R^1$ and L quadratic, we find $\alpha^\dagger(S) = (t^2-1)/(t^2+1/(n-1))$, where $t \doteq n^{\frac{1}{2}}\bar{y}/s$. With the modification $\mathscr{A} = [0, \infty)$, we get $\alpha^\dagger(S) = (t^2-1)^+/(t^2+1/(n-1))$ and

$$\hat{y}(\alpha^\dagger(S), S) = \frac{(t^2-1)^+}{(t^2+1/(n-1))}\bar{y}. \tag{3.1}$$

Note that (3.1) is a "shrinker", that is, the estimate of location is moved from $\bar{y}$ towards 0 by a factor that becomes more important as t^2 becomes smaller, while the magnitude of the shift tends to 0 as $t^2 \to \infty$.

(c) With $\hat{y}(\alpha, S) = \alpha_1 + \alpha_2\bar{y}$, $\mathscr{A} = R^2$ and L quadratic, we find $\alpha_1^\dagger(S) = n\bar{y}$ and $\alpha_2^\dagger(S) = -(n-1)$ giving $\hat{y}(\alpha^\dagger(S), S) = \bar{y}$!

(d) Illustrations (a), (b) and (c) suggest by their innocuousness that cross-validation can do no (serious) wrong. However their statistical interest must be minor because, in each case, $\hat{y}(\alpha, S)$ is made to depend on S only through $\bar{y}$. How does cross-validation perform when we permit greater realism by allowing $\hat{y}(\alpha, S)$ to depend on the order statistic $y_{(1)} \leqslant \ldots \leqslant y_{(n)}$ of S? For simplicity of illustration we will consider only the following prescription, in which m is the minimum integer greater than or equal to $n/4$:

$$\hat{y}(\alpha, S) = \alpha \times (\text{average of the } 2m \text{ outermost } y_{(i)} \text{ values}) + (1-\alpha) \times (\text{average of the rest}) \tag{3.2}$$

with $\mathscr{A} = \{0\,(0{\cdot}1)\,1{\cdot}0\}$. We may note that this prescription has an inbuilt symmetry.

The results of a Monte Carlo study with $n = 7$ are summarized in Fig. 2 and in Table 1. Random samples were generated from three contrasting distributions, each symmetrical about 0:

(i) "Uniform", closely approximating a uniform distribution on the interval $(-0{\cdot}5, 0{\cdot}5)$;

(ii) "Normal", closely approximating a standard normal distribution;

(iii) "Cauchy", closely approximating a standard Cauchy distribution.

The study employed 3,000 samples for (i) and (iii), 2,000 for (ii). Since no simplification attaches to quadratic L, the empirical distribution functions of $|\hat{y}(\alpha^\dagger(S), S)|$ were calculated for $L = |y - \hat{y}|$ and are shown in Fig. 2. For this L, we call $\hat{y}(\alpha^\dagger(S), S)$ CROSS. For comparison, three other estimators were calculated:

MIDRANGE $= \frac{1}{2}(y_{(1)} + y_{(7)})$, optimal for the uniform case;

MEAN $= \bar{y}$, optimal for the normal case;

"MEDIAN" $= 0{\cdot}0592(y_{(3)} + y_{(5)}) + 0{\cdot}8816y_{(4)}$, optimal for the Cauchy case (see Barnett, 1966). As our terminology suggests, "MEDIAN" is practically indistinguishable in behaviour from the median $y_{(4)}$.

Perusal of Fig. 2 shows that while $\hat{y}(\alpha^\dagger(S), S)$ does not shine for the uniform case, it is up with the frontrunner in the other two cases, one of which is realistically heavy-tailed. Table 1 gives efficiencies based on average values of $|\text{estimate}|^p$ for

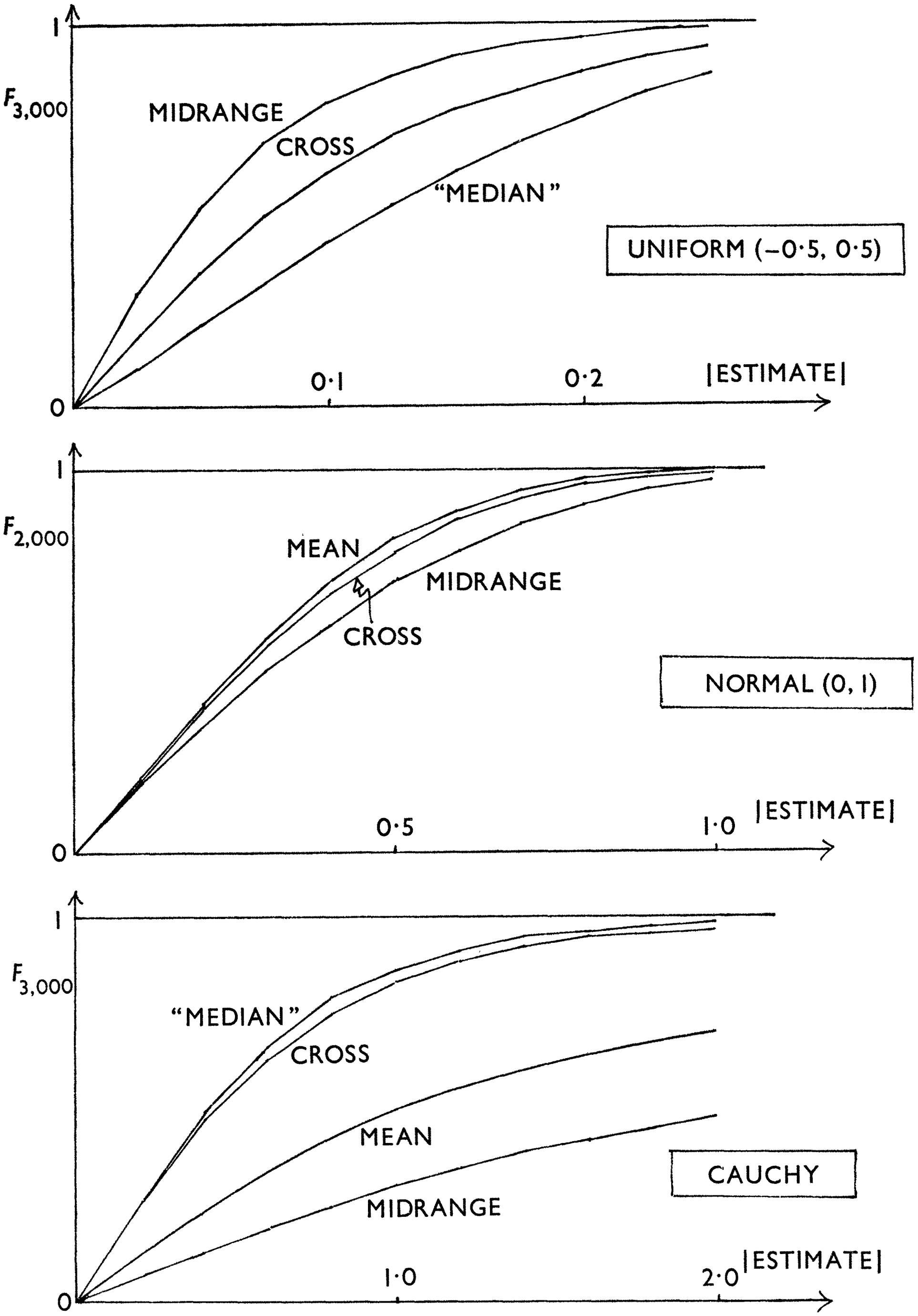

FIG. 2. Empirical distribution functions for the cross-validatory and comparison estimators; population mean 0 and $n = 7$.

$p = \frac{1}{2}, 1, 2$ against the best of MIDRANGE, MEAN, "MEDIAN". The two columns flanking CROSS give the efficiencies for cross-validatory choices using prescription (3.2) with $L = |y-\hat{y}|^{\frac{1}{2}}$ and $L = (y-\hat{y})^2$ in place of the modulus. Table 1 confirms the message of the distribution functions. However, the "experiment" with the alternative L's shows that it is not at all the rule that the best choice of power for L is the same as that used for the efficiency measure.

TABLE 1

Efficiencies for the Monte Carlo study

					Cross-validatory	*choice*	
Distribution	p	MIDRANGE	MEAN	"MEDIAN"	$\lvert y-\hat{y}\rvert^{\frac{1}{2}}$	CROSS	$(y-\hat{y})^2$
	$\frac{1}{2}$	1	0·82	0·66	0·80	0·79	0·81
Uniform	1		0·71	0·47	0·66	0·64	0·67
	2		0·58	0·26	0·46	0·43	0·50
	$\frac{1}{2}$	0·90	1	0·94	0·97	0·95	0·90
Normal	1	0·80		0·88	0·97	0·94	0·88
	2	0·62		0·77	0·98	0·95	0·90
	$\frac{1}{2}$	0·03	0·05	1	0·92	0·95	0·88
Cauchy	1				0·80	0·86	0·69
	2				0·4	0·5	

Further studies would be useful to assess the performance of CROSS relative to other robust estimators such as Hampel's M estimator and others investigated in Andrews *et al.* (1972).

The remainder of our examples are concerned with the more general problem of prediction of location in the presence of concomitant information.

Example 3.2. *The k-group problem.* Here we suppose that S consists of r real-valued observations in each of k groups:

$$S = \{(i, y_{ij}) \mid i = 1, \ldots, k;\, j = 1, \ldots, r\}.$$

Here x provides the group identification and $n = kr$. With the prescription

$$\hat{y}(i;\ \alpha, S) = \alpha \times (\text{overall average}) + (1-\alpha) \times (\text{average of } i\text{th group}), \tag{3.3}$$

$\mathscr{A} = R^1$ and quadratic L, we obtain by straightforward calculation

$$\alpha^{\dagger}(S) = \frac{n-1}{k(r-1)F+k-1}, \tag{3.4}$$

where $F = MS_b/MS_w$ is the customary F-ratio. Then

$$\hat{y}(i;\ \alpha^{\dagger}(S), S) = \bar{y} + (1-\alpha^{\dagger}(S))(\bar{y}_i - \bar{y}), \tag{3.5}$$

where $\bar{y} = \Sigma \bar{y}_i/k$. With the restriction $\mathscr{A} = [0,1]$, the multiplier of $\bar{y}_i - \bar{y}$ in (3.5) becomes $(1-\alpha^{\dagger}(S))^{+}$. The predictor (3.5) in either its modified or unmodified form may be compared with other estimators that share the property of shrinkage towards the overall mean, such as Lindley's adaptation (1962) of the James–Stein estimator.

Among such comparisons, Geisser (1974) has discovered by numerical study that the modified form is remarkably close to the Lindley "modal" estimator, although it is much easier to calculate. One interesting difference between the cross-validatory estimator and the Lindley–James–Stein estimator is that the former becomes operative for $k \geqslant 2$ while the latter does not shrink unless $k \geqslant 4$.

The predictor (3.5) finds immediate application to the Baseball Batting Average data analysed by Efron and Morris (1973). For this application, $k = 14$, the number of baseball players, $r = 45$, the number of battings of each player, while $y_{ij} = 0$ or 1 according as the ith player did not or did make a "hit" on his jth batting. We find $F = 1{\cdot}143$ and $\alpha^{\dagger}(S) = 0{\cdot}877$. The sum of the squares of the errors of the predictions of the batting averages for the remainder of the season is 0·0121 for our predictor, compared with 0·0146 for the Lindley–James–Stein predictor and with 0·0120 for the predictor $\bar{y}$ sceptically proposed by Plackett in the discussion of the Efron and Morris paper.

A cross-validatory analysis of Example 3.2 was outlined in the discussion of Efron and Morris (1973) using a prescription differing slightly from (3.3). The two associated estimators have similar properties. Prescription (3.3) is to be preferred here because it leads to a useful generalization: "model-mix".

Example 3.3. *Uniform flattening in multiple regression.* Suppose $y \in R^1$ and $x = (x(1), \ldots, x(p)) \in R^p$ and we take the prescription

$$\hat{y}(x;\, \alpha, S) = \alpha\bar{y} + (1-\alpha)\{\bar{y} + \Sigma b_k(x(k) - \bar{x}(k))\} \tag{3.7}$$

with $\mathscr{A} = R^1$ and quadratic L, where $\bar{x}(k) = \Sigma_j x_j(k)/n$ and b_k is the regression coefficient of y on $x(k)$ in the least-squares fitted multiple regression of y on x. We suppose that the usual rank condition is satisfied, requiring $n \geqslant p+1$. Relatively straightforward algebra yields

$$\alpha^{\dagger}(S) = \sum_{i=1}^{n} \left[\frac{r_i^2}{(1-A_{ii})^2} - \frac{nr_i(y_i-\bar{y})}{(n-1)(1-A_{ii})}\right] \Big/ \sum_{i=1}^{n} \left[\frac{r_i}{(1-A_{ii})} - \frac{n(y_i-\bar{y})}{(n-1)}\right]^2, \tag{3.8}$$

where r_i is the ith residual in the least-squares multiple regression and $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1(1) & \ldots & x_1(p) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n(1) & \ldots & x_n(p) \end{pmatrix}.$$

In the special case of equality of the A_{ii} (see, for example, Example 3.6 below), since $\operatorname{trace} \mathbf{A} = p+1$, we have $A_{ii} = (p+1)/n$ and then

$$\alpha^{\dagger}(S) = \frac{p}{(n-2p-1)}\left(\frac{1-R^2}{R^2 + [p^2/(n-1)(n-2p-1)]}\right), \tag{3.9}$$

where R^2 is the sample multiple correlation coefficient of y and x. With (3.9), the predictor becomes

$$\bar{y} + \sum_{k=1}^{p} f(R^2)\, b_k(x(k) - \bar{x}(k)),$$

where $f(R^2) = 1 - \alpha^{\dagger}(S)$. That is, the sample regression coefficients $b_1, \ldots, b_n$ are uniformly flattened by a multiplier

$$1 - \frac{p}{(n-2p-1)}\left(\frac{1-R^2}{R^2 + [p^2/(n-1)(n-2p-1)]}\right). \tag{3.10}$$

The expression (3.10) should be compared with a class of flatteners suggested by Stein (1960) and, in particular, with the subclass given by Baranchik (1973) for the case $p \geqslant 3$ and $n \geqslant p+2$:

$$f_c(R^2) = 1 - c\left(\frac{1-R^2}{R^2}\right), \tag{3.11}$$

where $0 < c < 2(p-2)/(n-p+2)$. Very roughly, for p moderate and n even larger, (3.10) corresponds to choosing c at the midpoint of Baranchik's range of values. The term "flattener" is strictly justified if we take $\mathscr{A} = [0, 1]$.

Example 3.4. *Choice of variables for least-squares prediction.* As a development intended to be considered alongside the method exemplified so far, we introduce the combination of two-stage cross-validatory choice ("double-cross') and *linear prescription.* By the latter, we mean linearity of $\hat{y}$ with respect to some known functions of the concomitant variable x. We express this linearity in a way that will prepare the ground for double-cross:

$$\hat{y}(x; \alpha) = \sum_{k=1}^{p} a_k b_k c_k(x), \tag{3.12}$$

where the $c_k(.)$ are specified functions and the elements of $\mathbf{a} = (a_1, \ldots, a_p)$ are either 0 or 1, corresponding to a "choice of the variables". For $\mathbf{a}$ such that only $a_{k_1}, \ldots, a_{k_q}$, with $k_1 < \ldots < k_q$, are non-zero, write $\mathbf{b} = (b_{k_1}, \ldots, b_{k_q}) \in R^q$. Thus, in the notation of Section 2, VI, $B(\mathbf{a}) = R^q$ while A is some set of "choices of the variables". The possibility of allowing $p = \infty$ should be recognized; but the challenge raised by this possibility will not be taken up in this paper. Instead we will put the customary, if unscientific, restriction on q to ensure non-singularity. We take L to be quadratic.

Fixing $\mathbf{a}$, two cases are to be distinguished. Write $\mathbf{c}_k = (c_k(x_1), \ldots, c_k(x_n))'$ and $\mathbf{X}(\mathbf{a}) = (\mathbf{c}_{k_1} \ldots \mathbf{c}_{k_q})$.

Case 1: $n \geqslant q+2$ and all $(n-2) \times q$ submatrices of $\mathbf{X}(\mathbf{a})$ are of full rank q.

Case 2: The conditions of Case 1 do not obtain.

In Case 2, for at least one (i,j), $\mathbf{b}^\dagger(S_{\backslash ij}, \mathbf{a})$ is not uniquely defined. (In fact, for such (i,j) the average of L over the sample $S_{\backslash ij}$ can be made *zero* for infinitely many values of $\mathbf{b}$.) So, in Case 2, $C^{\dagger\dagger}$ could not be calculated. In Case 1, $\mathbf{b}(S_{\backslash i}, \mathbf{a})$ is the least-squares estimate of $\mathbf{b}$, $[\mathbf{X}_{\backslash i}(\mathbf{a})^{\mathrm{T}} \mathbf{X}_{\backslash i}(\mathbf{a})]^{-1} \mathbf{X}_{\backslash i}(\mathbf{a})^{\mathrm{T}} \mathbf{y}_{\backslash i}$ where $\mathbf{X}_{\backslash i}(\mathbf{a})$ denotes the matrix $\mathbf{X}(\mathbf{a})$ with its ith row omitted and $\mathbf{y}_{\backslash i}$ is similiarly defined.

So we consider only those values of $\mathbf{a}$ for which Case 1 obtains. For such $\mathbf{a}$, we find that

$$C^\dagger(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} \frac{[y_i - \hat{y}_i(\mathbf{a})]^2}{(1 - A_{ii}(\mathbf{a}))^2}, \tag{3.13}$$

where $\hat{y}_i(\mathbf{a})$ is, for fixed $\mathbf{a}$, the least-squares fitted value for y_i in a conventional least-squares fit of the prescription (3.12) to the sample S and where

$$\mathbf{A}(\mathbf{a}) = \mathbf{X}(\mathbf{a}) [\mathbf{X}(\mathbf{a})^{\mathrm{T}} \mathbf{X}(\mathbf{a})]^{-1} \mathbf{X}(\mathbf{a})^{\mathrm{T}}, \tag{3.14}$$

the projection matrix for the least-squares procedure.

The quantity (3.13) has been considered by Hocking (1972) in a way that may be considered to parallel the work of the present section.

It may be supposed that, in any application, A is a natural manageable set of $\mathbf{a}$ values all falling into Case 1. Then $\mathbf{a}^{\dagger\dagger}(S)$ will be the $\mathbf{a}$ minimizing (3.13) and $\mathbf{b}^{\dagger\dagger}(S)$ will be the "least-squares estimate" of $\mathbf{b}$,

$$[\mathbf{X}(\mathbf{a}^{\dagger\dagger}(S))^{\mathrm{T}}\mathbf{X}(\mathbf{a}^{\dagger\dagger}(S))]^{-1}\mathbf{X}(\mathbf{a}^{\dagger\dagger}(S))^{\mathrm{T}}\mathbf{y}.$$

To calculate the cross-validatory assessment measure $C^{\dagger\dagger}$ for $\alpha^{\dagger\dagger}(S)$ we need $\mathbf{a}^{\dagger\dagger}(S_{\backslash i})$ for each i. With $S_{\backslash i}$ replacing S, it is readily shown that the equivalent of (3.13) is

$$\frac{1}{(n-1)}\sum_{\backslash i}\frac{[y_j-\hat{y}_j(\mathbf{a})+A_{ji}(\mathbf{a})\,(y_i-\hat{y}_i(\mathbf{a}))/(1-A_{ii}(\mathbf{a}))]^2}{[1-A_{jj}(\mathbf{a})-A_{ji}^2(\mathbf{a})/(1-A_{ii}(\mathbf{a}))]^2}, \tag{3.15}$$

which is therefore readily calculated from the ingredients of the calculation of (3.13). Then $\mathbf{a}^{\dagger\dagger}(S_{\backslash i})$ is the value of $\mathbf{a}$ minimizing (3.15) and we finally obtain

$$C^{\dagger\dagger}=\frac{1}{n}\sum_{i=1}^{n}\frac{[y_i-\hat{y}_i(\mathbf{a}^{\dagger\dagger}(S_{\backslash i}))]^2}{[1-A_{ii}(\mathbf{a}^{\dagger\dagger}(S_{\backslash i}))]^2}. \tag{3.16}$$

For the case when A is finite and hierarchical with respect to the "variables", $c_1(.), c_2(.), \ldots$, that is, with the equivalence $\mathbf{a}\sim(k_1,\ldots,k_q)$, when

$$A=\{(1),(1,2),\ldots,(1,2,\ldots,K)\} \tag{3.17}$$

for some K, the computer program DOUBLECROSS has been devised to accomplish the necessary calculations.

Application to the geopotential determined by 27 *Earth satellites*

We apply DOUBLECROSS in re-analysis of satellite data collected and analysed by King-Hele and Cook (1973). Quoting from King-Hele *et al.* (1969):

> "... the Earth's gravitational potential U at an exterior point distant r from the Earth's centre, and having geocentric latitude ϕ, is written in a series of spherical harmonics as
>
> $$U=\frac{GM}{r}\left\{1-\sum_{n=2}^{\infty}J_n\left(\frac{R}{r}\right)^n P_n(\sin\phi)\right\} \tag{U}$$
>
> where G is the gravitational constant, M the mass of the Earth, and R the Earth's equatorial radius. $P_n(\sin\phi)$ is the Legendre polynomial of degree n and argument $\sin\phi$, and J_n are constant coefficients. Equation (U) does not take into account the small variation of U with longitude and represents an average over all longitudes."

The orbit of an Earth satellite is nearly elliptical with an eccentricity of the form $\alpha+\beta\sin\omega$, where ω is the angle, measured round the orbit, between the northward equator crossing and the point of nearest approach to the Earth. The term β is a known linear function of the unknown odd-order Legendre coefficients $J_3, J_5, \ldots$. An estimate of β can be made from repeated tracking observations of the satellite from one or more ground stations. Thus each satellite produces a single derived observation which is linearly related to $J_3, J_5, \ldots$. King-Hele and Cook fitted the equation

$$Y=10^6(F_3J_3+F_5J_5+\ldots) \tag{3.18}$$

to the data which are reproduced in Table 2.

TABLE 2, VALUES OF F_3, F_5, F_7,...,F_{33}, Y AND σ FOR THE 27 SATELLITES

Satellite	F_3	F_5	F_7	F_9	F_{11}	F_{13}	F_{15}	F_{17}	F_{19}	F_{21}	F_{23}	F_{25}	F_{27}	F_{29}	F_{31}	F_{33}	Y	σ
Explorer 11	-1	0.8524	-0.0540	-0.4925	0.4513	-0.0850	-0.1985	0.2173	-0.0627	-0.0793	0.1054	-0.0403	-0.0314	0.0520	-0.0246	-0.0119	2.424	0.02
L.C.S.1	-1	0.4116	0.1129	-0.1624	0.0437	0.0210	-0.0197	0.0034	0.0033	-0.0022	0.0002	0.0005	-0.0002	-0.0000	0.0001	-0.0000	2.351	0.1
OSO 3	-1	0.6597	0.4257	-0.8663	0.3095	0.4292	-0.5376	0.0666	0.3463	-0.2930	-0.0453	0.2443	-0.1387	-0.0793	0.1557	-0.0517	2.292	0.05
Vanguard 2	-1	0.4949	0.2476	-0.3963	0.1135	0.1309	-0.1367	0.0139	0.0648	-0.0475	-0.0068	0.0312	-0.0157	-0.0083	0.0146	-0.0043	2.304	0.05
Explorer 27	-1	-0.0952	0.7329	-0.1079	-0.3926	0.1406	0.1856	-0.1132	-0.0774	0.0760	0.0265	-0.0459	-0.0052	0.0256	-0.0022	-0.0132	2.272	0.03
Telstar 1	-1	-0.3060	0.3865	0.0973	-0.1404	-0.0342	0.0553	0.0132	-0.0234	-0.0054	0.0104	0.0023	-0.0048	-0.0010	0.0023	0.0005	2.460	0.02
Echo 1 r	-1	-0.6074	0.5101	0.3298	-0.1958	-0.1610	0.0679	0.0745	-0.0212	-0.0332	0.0056	0.0144	-0.0009	-0.0060	-0.0002	0.0025	2.407	0.05
Anna 1B	-1	-1.0319	0.4520	0.7293	-0.0691	-0.4176	-0.0695	0.2058	0.0898	-0.0861	-0.0696	0.0275	0.0437	-0.0033	-0.0238	-0.0044	2.561	0.03
Ariel 2	-1	-1.3770	0.3627	1.1538	0.1742	-0.7363	-0.3979	0.3586	0.4124	-0.0924	-0.3244	-0.0597	0.2069	0.1228	-0.1007	-0.1278	2.721	0.05
Tiros 5	-1	-3.1912	-1.5028	1.2261	1.9106	0.5435	-0.8739	-0.9768	-0.1240	0.5467	0.4684	-0.0220	-0.3190	-0.2122	0.0560	0.1773	4.177	0.05
Explorer 29	-1	-3.2468	-1.5575	0.5220	1.0215	0.4162	-0.1811	-0.2953	-0.1084	0.0603	0.0861	0.0285	-0.0200	-0.0256	-0.0075	0.0067	4.051	0.1
Explorer 32	-1	9.2127	9.9075	3.9702	-2.1581	-4.5901	-3.2564	-0.3820	1.7142	2.0178	0.9509	-0.3728	-1.0458	-0.8485	-0.1683	0.4226	-2.892	0.03
Transit 4A	-1	2.8256	3.9667	2.4636	0.1747	-1.3105	-1.4968	-0.8076	0.0294	0.5055	0.5156	0.2477	-0.0401	-0.1866	-0.1729	-0.0732	0.803	0.03
Secor 5	-1	0.8959	1.4320	1.0040	0.3610	-0.0753	-0.2258	-0.1881	-0.0877	-0.0055	0.0324	0.0343	0.0199	0.0047	-0.0041	-0.0062	1.857	0.05
FR-1	-1	-0.3437	0.3420	0.7567	0.8615	0.7294	0.4716	0.1925	-0.0347	-0.1747	-0.2254	-0.2062	-0.1463	-0.0739	-0.0101	0.0333	2.371	0.05
Alouette 2	-1	-0.5688	-0.1803	0.0578	0.1721	0.2045	0.1898	0.1527	0.1091	0.0681	0.0343	0.0095	-0.0067	-0.0155	-0.0188	-0.0182	2.736	0.1
Explorer 20	-1	-0.6619	-0.2437	0.0735	0.2604	0.3346	0.3287	0.2759	0.2033	0.1301	0.0676	0.0209	-0.0096	-0.0259	-0.0316	-0.0301	2.803	0.1
Essa 1	-1	-0.8164	-0.5015	0.2124	0.0086	0.1554	0.2363	0.2651	0.2571	0.2260	0.1831	0.1371	0.0937	0.0565	0.0269	0.0053	2.903	0.1
Midas 4	-1	-0.4592	-0.1740	-0.0583	-0.0167	-0.0034	0.0001	0.0007	0.0005	0.0003	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	2.697	0.02
1963-49B	-1	0.9106	-0.7255	-0.5549	-0.4169	-0.3102	-0.2295	-0.1692	-0.1244	-0.0913	-0.0669	-0.0490	-0.0359	-0.0262	-0.0192	-0.0140	3.070	0.02
Explorer 42	-1	2.1228	-3.1454	3.9835	-4.6112	5.0320	-5.2642	5.3335	-5.2681	5.0958	-4.8427	4.5320	-4.1838	3.8151	-3.4397	3.0685	0.490	0.62
Dial	-1	1.8904	-2.4974	2.8223	-2.9160	2.8387	-2.6456	2.3820	-2.0826	1.7730	-1.4712	1.1890	-0.9336	0.7088	-0.5158	0.3539	1.190	0.22
Peole	-1	1.7583	-1.9740	1.6766	-1.0595	0.3592	0.2277	-0.5880	0.6978	-0.6035	0.3886	-0.1425	-0.0646	0.1930	-0.2345	0.2051	2.510	0.05
Cosmos 248	-1	-13.3354	-12.8704	-1.6928	8.6386	9.8900	3.1156	-4.4779	-6.7651	-3.2588	1.8992	4.3105	2.7973	-0.4913	-2.5750	-2.1522	11.340	0.05
Cosmos 44	-1	7.4360	9.1660	4.3817	-1.7672	-4.8682	-3.9373	-0.8815	1.6943	2.3593	1.3497	-0.1449	-1.0414	-0.9935	-0.3531	0.2847	-2.106	0.05
Geos 2	-1	-0.1071	0.5041	0.6763	0.5585	-0.3312	0.1195	-0.0201	-0.0823	-0.0880	-0.0645	-0.0338	-0.0088	0.0060	0.0114	0.0107	2.359	0.02
Prospero	-1	-017615	-0.4375	-0.1728	0.0098	0.1187	0.1706	0.1829	0.1703	0.1444	0.1132	0.0823	0.0547	0.0320	0.0147	0.0024	2.910	0.1

Their fitting was based on a weighted least-squares procedure using the standard deviations σ which were believed to provide a realistic assessment of the accuracy of the 27 individual observations. After some delicate comparisons involving alternative choices of the "variables", King-Hele and Cook recommended the estimate of $(J_3, J_5, \ldots)$

$$\left. \begin{array}{l} J_{2k+1} \hat{=} 0, \\ \begin{pmatrix} J_3 \\ J_5 \\ J_7 \\ J_9 \\ J_{11} \\ J_{13} \\ J_{15} \\ J_{17} \end{pmatrix} \hat{=} \begin{pmatrix} -2{,}531 \\ -246 \\ -326 \\ -94 \\ 159 \\ -131 \\ -26 \\ -258 \end{pmatrix} \times 10^{-9} \end{array} \right\}, \qquad (3.19)$$

the latter values being the weighted least-squares estimates for a fit of the first eight "variables". The estimate (3.19) is intended to be combined with an estimate of $(J_2, J_4, J_6, \ldots)$ from other satellite data to give an estimate of U.

Translating into present terminology, we have $n = 27$, $p = \infty$, $y_i = Y_i/\sigma_i$, $x_i =$ "other data for ith satellite", $c_k(x_i) = F_{2k+1}(i\text{th satellite})/\sigma_i$ and $b_k = 10^6 J_{2k+1}$. We will use (3.17) with $K = 16$ rather than $K = 25$, the maximum possible value consistent with Case 1, since J_{33} was the highest order coefficient considered by King-Hele and Cook. Using q as an index for $(1, 2, \ldots, q) \in A$, Fig. 3 shows the behaviour of $C^\dagger(q)$ and, for comparison purposes, that of

$$\mathrm{MSE}_q \overset{\text{def}}{=} \sum_{i=1}^{n} [y_i - \hat{y}_i(q)]^2/(n-q).$$

We see that $\mathbf{a}^{\dagger\dagger}(S) = (1, 2, \ldots, 8)$ so that the cross-validatory estimate of the coefficients is also (3.19). DOUBLECROSS gives us $C^{\dagger\dagger} = 1{\cdot}63$ which may be compared with $C^\dagger(8) = 1{\cdot}17$. The difference between the two values provides a necessary correction for the "selection of variables effect" within our selected hierarchy A. An idea of the robustness of the choice $\mathbf{a}^{\dagger\dagger}(S)$ is also provided by other output of DOUBLECROSS which informs us that $\mathbf{a}^{\dagger\dagger}(S_{\setminus i})$ is $(1, 2, \ldots, 8)$ for all i except 22 (satellite Dial) and 23 (satellite Peole) when it is $(1, 2, \ldots, 9)$.

It is impossible to apply the general method to King-Hele and Cook's choice since this was unformalized. However, were it possible to transform King-Hele and Cook into a memoryless machine that would slavishly follow their mental processes, we could obtain a cross-validatory assessment measure $C^\dagger$ by feeding the machine with 27 data sets, each with one satellite omitted, and this value of $C^\dagger$ could be compared with our $C^{\dagger\dagger}$ of 1·63.

Example 3.5. *The model-mix prescription.* Examples 3.2 and 3.3 contain a useful signpost towards generality. To be specific, prescription (3.7) may be viewed as a weighted mixture of two prescriptions

$$\hat{y}(x;\, \alpha, S) \equiv \bar{y},$$

arguably relevant for the "model"

$$M_1\text{: No relationship of } y \text{ to } x,$$

and

$$\hat{y}(x;\, \alpha, S) \equiv \bar{y} + \sum_{k=1}^{p} b_k(x(k) - \bar{x}(k)),$$

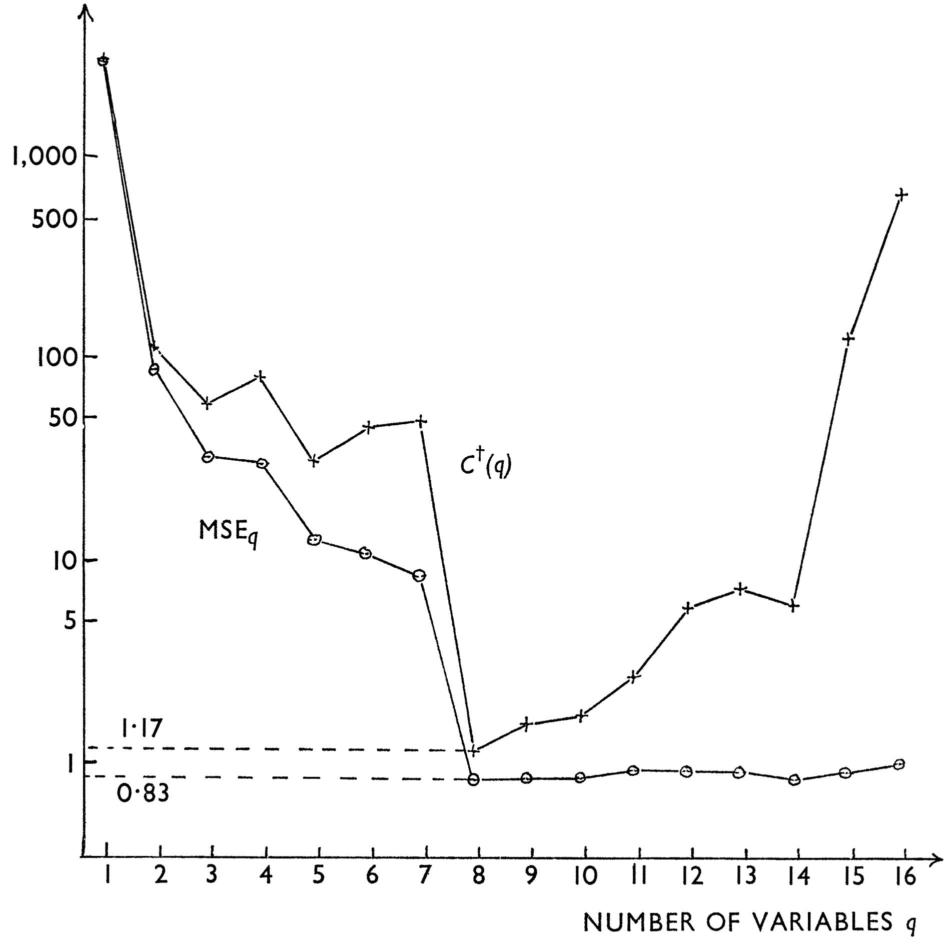

FIG. 3. DOUBLECROSS analysis of satellite data.

arguably relevant for the "model"

$$M_2\text{: Strong dependence of } y \text{ on each of } x(1), \ldots, x(p).$$

A weakness of (3.7) may be considered to lie in the polarity or non-exhaustiveness of M_1 and M_2.

Generally, suppose there is sufficient structure to suggest a set of indicative, roughly exhaustive models, $M_1, \ldots, M_m$, and that, for each M_q, there is a natural, reasonable predictor $\hat{y}^{(q)}(x, S)$ say. Then the *model-mix prescription* is

$$\hat{y}(x;\, \alpha, S) = \sum_{q=1}^{m} \alpha_q \hat{y}^{(q)}(x, S)$$

with suitable $\mathscr{A}$. With $\mathscr{A} = R^m$ and L quadratic, we get

$$\boldsymbol{\alpha}^{\dagger}(S) = (\alpha_1^{\dagger}(S), \ldots, \alpha_m^{\dagger}(S))^{\mathrm{T}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}, \tag{3.20}$$

where

$$\mathbf{Z} = \begin{pmatrix} \hat{y}^{(1)}(x_1, S_{\backslash 1}) & \cdots & \hat{y}^{(m)}(x_1, S_{\backslash 1}) \\ \vdots & \vdots & \vdots \\ \hat{y}^{(1)}(x_n, S_{\backslash n}) & \cdots & \hat{y}^{(m)}(x_n, S_{\backslash n}) \end{pmatrix},$$

assumed to be of full rank.

In the case where, for each q, M_q is a linear model and $\hat{y}^{(q)}(x, S)$ the corresponding least squares predictor, the calculation of $\mathbf{Z}$ is facilitated by the identity

$$\hat{y}^{(q)}(x_i, S_{\backslash i}) = \frac{\hat{y}^{(q)}(x_i, S)}{(1 - A_{ii}^{(q)})} - \frac{A_{ii}^{(q)}\, y_i}{(1 - A_{ii}^{(q)})},$$

where $\mathbf{A}^{(q)}$ is the projection matrix associated with the least-squares procedure for the model M_q (cf. (3.14)).

Application to the satellite data

We will take the hierarchical models

M_q: Only the first q Legendre polynomials have
non-zero coefficients.

Computer program MODELMIX mixes least-squares predictors and, for a selected value of m, calculates $\boldsymbol{\alpha}^{\dagger}(S)$, the corresponding estimates of the Legendre coefficients and the assessment measure $C^{\dagger}$.

For the choice $m = 9$, the values are

$$\boldsymbol{\alpha}^{\dagger}(S) = \begin{pmatrix} 0{\cdot}0040 \\ -0{\cdot}0233 \\ 0{\cdot}0377 \\ -0{\cdot}0312 \\ -0{\cdot}0078 \\ 0{\cdot}1005 \\ -0{\cdot}0874 \\ 0{\cdot}5277 \\ 0{\cdot}4820 \end{pmatrix}, \quad \begin{pmatrix} J_3 \\ J_5 \\ J_7 \\ J_9 \\ J_{11} \\ J_{13} \\ J_{15} \\ J_{17} \\ J_{19} \end{pmatrix} \triangleq \begin{pmatrix} -2{,}540 \\ -235 \\ -337 \\ -91 \\ 168 \\ -152 \\ -51 \\ -255 \\ -16 \end{pmatrix} \times 10^{-9}, \quad C^{\dagger} = 0{\cdot}875. \tag{3.21}$$

We remark that (i) the components of $\boldsymbol{\alpha}^{\dagger}(S)$ sum to 1·002 which is close to unity, (ii) $\boldsymbol{\alpha}^{\dagger}(S)$ puts most weight on M_8 and M_9 and (iii) the value of $C^{\dagger}$ is about half the value of $C^{\dagger\dagger}$ for DOUBLECROSS. However, the behaviour of MODELMIX for other values of m remains to be analysed.

It would be of interest to investigate the effect of the natural restriction of $\mathscr{A}$ to the simplex

$$\mathscr{A} = \{\boldsymbol{\alpha} \,|\, \alpha_q \geqslant 0,\ q = 1, \ldots, m,\ \Sigma \alpha_q = 1\},$$

using quadratic programming techniques.

It is worth noting that the application of DOUBLECROSS in Example 3.4 is more economically considered as an application of MODELMIX with $m = 16$ and

$$\mathscr{A} = \{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)\}.$$

With this change, $C^\dagger(q) \to C(\alpha)$ and $C^{\dagger\dagger} \to C^\dagger$.

Example 3.6. *Cases with symmetries*; *standard experimental designs.* In Example 3.3 we postulated, in somewhat arbitrary fashion, a special case of equality of the A_{ii}. Inspection of the formula for $C^\dagger(\mathbf{a})$ in Example 3.4 shows that the same special case there yields

$$C^\dagger(\mathbf{a}) = \frac{n}{(n-q)} \text{MSE}(\mathbf{a}), \tag{3.22}$$

where MSE(**a**) denotes the customary mean-square error for a least-squares fit of the linear prescription indexed by **a**. When the variation of $C^\dagger(\mathbf{a})$ with **a** is examined, we see that, since n is constant, we are equivalently required to divide the residual sum of squares by *the square of the degrees of freedom.* This requirement is so strikingly different from standard methods that some further illustration of its possible value is necessary.

The simplest illustrations are provided by the analysis of standard balanced experimental designs. To think about these, it is preferable to avoid the notation of (3.12). Rather we consider prescriptions of the form

$$\begin{aligned} y(x_i;\, \alpha) = y(x_i;\, (\mathbf{a}, \mathbf{b})) = b_0 + & \\ & + b_{1i_1} + b_{1i_2} + \ldots \\ & + b_{12i_1i_2} + \ldots \\ & + \ldots, \end{aligned} \tag{3.23}$$

where **a** controls the order of the expression on the right, $i = (i_1, i_2, \ldots)$ is the index in natural correspondence with the block structure and treatment allocation of the design and the components of **b** satisfy

$$\begin{aligned} \Sigma_{i_1} b_{1i_1} &= \Sigma_{i_2} b_{2i_2} &= \ldots = 0, \\ \Sigma_{i_1} b_{12i_1i_2} &= \Sigma_{i_2} b_{12i_1i_2} &= \ldots = 0, \\ &\vdots \end{aligned}$$

but are otherwise unrestricted. The final component of i may not appear on the right of (3.23) if it corresponds to replication. Such prescriptions correspond to symmetrical linear models up to a certain interaction order for standard balanced experimental designs.

If $\mathscr{S}$ denotes the vector space of $(\hat{y}(x_1;\, \alpha) \ldots \hat{y}(x_n;\, \alpha))^{\mathrm{T}}$ generated for some fixed **a** then A_{ii} has the simple representation-free interpretation as the cosine of the angle between $\mathscr{S}$ and the vector $(0 \ldots 0\ 1\ 0 \ldots 0)^{\mathrm{T}}$ along the ith axis. The symmetry of (3.23) with respect to permutations within $\{i_1\}, \{i_2\}, \ldots$ shows that the A_{ii} must be equal.

Application to the analysis of a 2^5 factorial experiment without replication

We will consider the "Pilot Plant" data given on pp. 183–186 of Johnson and Leone (1964). The standard analysis of variance is summarized in Table 3. We will take four obvious choices of $\mathbf{a}$ in the specialization of (3.23) and calculate the corresponding values of $C^{\dagger}(\mathbf{a})$. We have $i = (i_1, \ldots, i_5)$ for the levels of $A, \ldots, E$ respectively. The four values of $\mathbf{a}$ may be listed as follows:

TABLE 3

Summary of analysis of variance for Pilot Plant data

Sums of squares				
	AB 2·000	ABC: 6·125		
	AC: 3·125	ABD: 2·000		
	AD: 0·500	ABE: 1·125	$BCDE$	
A: 40·500**	AE: 3·125	ACD: 0·125	$ACDE$	
B: 0·500	BC: 10·125	ACE: 0·500	$ABDE$	16·250
C: 15·125	BD: 0·500	ADE: 28·125*	$ABCE$	
D: 40·500**	BE: 3·125	BCD: 10·125	$ABCD$	
E: 55·125**	CD: 3·125	BCE: 4·500	$ABCDE$	
	CE: 12·500	BDE: 1·125	**: $P<0{\cdot}01$	
	DE: 15·125	CDE: 0·500	*: $P<0{\cdot}05$	
$\sum$: 151·750	$\sum$: 53·250	$\sum$: 54·250	$\sum$: 16·250	

$\mathbf{a}^{(1)}$: $\hat{y}(x_i;\ \alpha) = b_0$ [no dependence on any factor],
$\mathbf{a}^{(2)}$: $\hat{y}(x_i;\ \alpha) =$ first 2 rows of (3.23) [additivity],
$\mathbf{a}^{(3)}$: $\hat{y}(x_i;\ \alpha) =$ first 3 rows of (3.23) [up to first-order interactions],
$\mathbf{a}^{(4)}$: $\hat{y}(x_i;\ \alpha) =$ first 4 rows of (3.23) [up to second-order interactions].

The $C^{\dagger}(\mathbf{a})$ values, with those of MSE for comparison, are

	$\mathbf{a}^{(1)}$	$\mathbf{a}^{(2)}$	$\mathbf{a}^{(3)}$	$\mathbf{a}^{(4)}$
$C^{\dagger}$ (a)	9·2	5·9	8·8	14·4
MSE(a)	8·9	4·8	4·4	2·7

We see here the same phenomenon as in Example 3.4; $C^{\dagger}(\mathbf{a})$ falls but then begins to increase revealing the penalty for "overprescribing".

Another illustration is provided by the "Two-variable Example" of Gorman and Toman (1966) in which, it may be verified, the condition of equality of the A_{ii} also obtains. Gorman and Toman used Mallows C_p which has an objective analogous to that of $C^{\dagger}(\mathbf{a})$. We can supplement one of their tables as follows:

Variables in equation	C_p	$C^{\dagger}(\mathbf{a})$
None	84·6	5·49
X_1	4·1	1·02
X_2	2·5	0·91
X_1, X_2	3·0	0·95

In this example, at least, C_p has a more dramatic minimum at X_2 than $C^{\dagger}(\mathbf{a})$ has. However, the similarity is of interest.

4. Review and Discussion

How does previous work on cross-validation relate to what we have done? Hills (1966) considered the problem of proper assessment of the error rates of a discriminant whose form is already fully specified. In our terminology, his application of cross-validation has a prescription of type $\hat{y}(x; S)$, with no α to be chosen, and the assessment is the calculation of $C^\dagger$. Hills includes a multinomial example, which we have not covered, but which is suggestive of application to the simple forms of medical diagnosis of Teather (1973). The same specialization is to be found in Kerridge's discussion of Hills (1966) and, in greater detail, in Lachenbruch and Mickey (1968). Mosteller and Tukey (1968) made a fascinating analysis of 22 Federalist papers. However, in essence, their cross-validation corresponds to a $\hat{y}(x; S)$ prescription based on jack-knifing.

This brief review informs us that previous work lacks the argument α in its prescriptions, which, as we have seen in our general framework and examples, can play an interesting role.

Turning towards generalities, Fig. 4 portrays a conventional paradigm for a large portion of statistical activity.

Linkage * in Fig. 4 represents the role of current data in model building; it may involve *tests of assumptions* about the model or the *selection* of the model from a restricted class. It may be present at a very informal level.

Linkage ** represents any attempt to assess the quality of the procedure by its *degree of fit* to the current data.

Linkage *** represents any follow-up on the application that may become available, its validation on future or additional data drawn from the area of applicability.

Rectangle ⊕ is alternative to ** and ***, differing from these in its direct dependence on the model.

A detailed and realistic analysis of almost any example will reveal that not all is well with the conventional paradigm. The principal difficulty lies in the complex interaction between activities in its different components. For example, in the problem of univariate estimation with data that arrives without a pedigree, it would be just too naive to test for normality and use the sample mean if the test is not significant (Mosteller and Tukey, 1968); but what *should* we do? Again, in the choice of variables in linear regression problems (where the true regression may be far from linear) are we to use F-tests to select the variables and, if so, in what order and at what significance level? The C_p criterion of Mallows and the MSEP measure of Allen (1971) and the J_p criterion in Hocking (1972) all have a basis in the conventional paradigm.

By contrast, the cross-validatory paradigm (Fig. 1) has a simpler structure and each of its parts has a relatively clearly defined role. There will almost certainly be major problems in the execution of the cross-validatory paradigm but the status of each problem that arises should not be as ambiguous as some that are associated with the conventional paradigm.

The Bayesian approach usually demands, and possibly deserves, to be given special consideration; its adherents are unlikely to accept the complexity of Fig. 4. Ignoring purely technical difficulties, the Bayesian paradigm has a stark simplicity:

$$\begin{aligned}&[\text{Determination of prior (which } \textit{includes} \text{ model specification!)}]\\ &\qquad\qquad \rightarrow [\text{Data analysis (Bayes theorem!)}] \rightarrow [\text{Application}]. \qquad (4.1)\end{aligned}$$

The alternative Bayesian paradigm

> [Data] → [Direct assignment of posterior] → [Coherence check that the assignment "coheres" with assignments that would have been made for data sets different from the one actually obtained] → [Application]

is potentially equally simple but would constitute an invitation to dishonesty. The purist Bayesian viewpoint is one that admits no element of choice in (4.1); you do not *choose* a prior, you *have* a prior. And, if there is no choice, there can be no criticism.

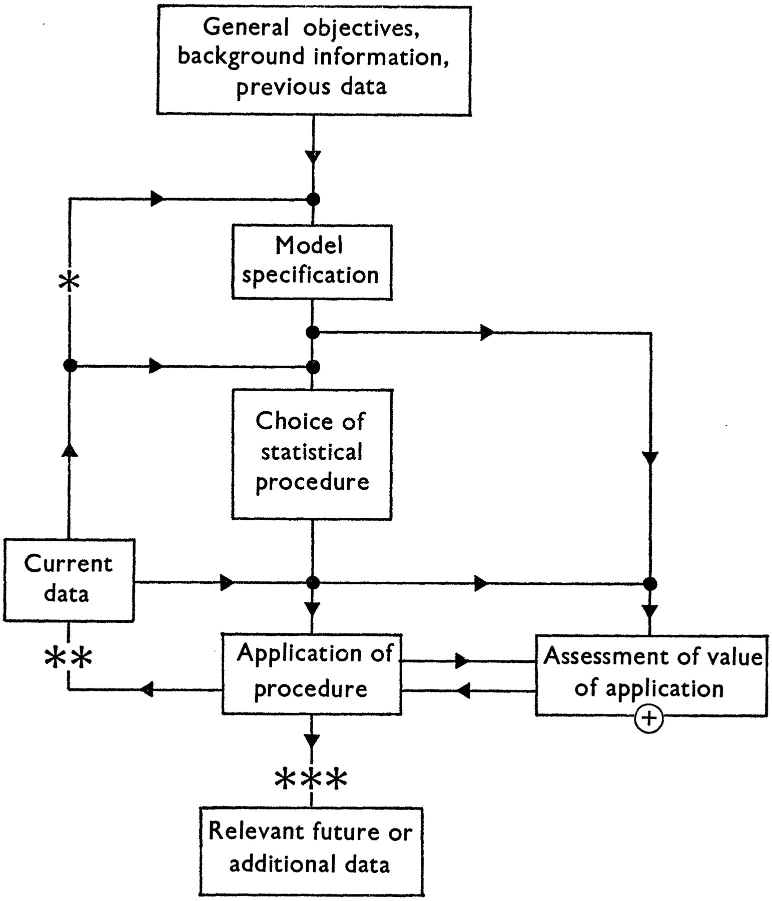
General objectives, background information, previous data
Model specification
*
Choice of statistical procedure
Current data
**
Application of procedure
Assessment of value of application
+
***
Relevant future or additional data

FIG. 4. Conventional paradigm.

The *contretemps*, revealed by Beale's puzzlement at the Bayesian endorsement of least-squares prediction in Lindley (1968), must be regarded as a consequence of this viewpoint; if the prior (of uniform, independent type) says that all parameters are likely to be large then one must take the consequences. However, Mosteller and Wallace (1963, pp. 306–307) write:

> "In the presence of large numbers of variables, preparing for selectivity is necessary—in classical studies through calibrating or validating sets of data whose origin is known but that are uncontaminated by the selection and weighting processes; in Bayesian studies through realistic priors that can be based on pools of variables."

So priors must be classified as realistic or unrealistic. Presumably a realistic prior is one that would not give its user any cross-validatory shocks of the kind described in Section 1. Recently progress has been made in developing priors representing "exchangeability" (Lindley and Smith, 1972); such priors are clearly supposed by their creators to be at least more realistic than earlier efforts. That realism may not be a very robust concept when applied to priors is illustrated by the very simple case of Example 3.1. With normality and a particular proper conjugate prior (De Groot, 1970, p. 170), the posterior mean is $\alpha\bar{y}$ where α, $0<\alpha<1$, is fixed by the prior. The realistic possibility of large $\bar{y}$ casts immediate doubt on the realism of this prior. While one must wish the recent Bayesian work every success, it is possible that the cross-validatory paradigm can offer a route that smoothly by-passes the Bayesian industrial area.

Finally, some quotations:

(i) "Many mathematical and philosophical discussions begin with a general theory from which are derived general principles; from these, in turn, specific procedures are produced and finally exemplified. In discussing data analysis, we find the following somewhat opposite order more practical. (a) First, what to do. (What treatment to apply to the data of a given sort of problem, arithmetically or graphically)." Mosteller and Tukey (1968, p. 81).

(ii) "The method studied by Lachenbruch, of allocating each observation in turn by a discriminant computed from the rest of the observations, has two important merits. It can be used in other problems whose solution is intractable and its rationale is readily understood by people who have little technical knowledge of statistics." Cochran, discussing Hills (1966).

(iii) "Obtaining a valid measure of uncertainty is not just a matter of looking up a formula." Mosteller and Tukey, (1968, p. 124).

(iv) "The word 'valid' would be better dropped from the statistical vocabulary. The only real validation of a statistical analysis, or of any scientific enquiry, is confirmation by independent observations." Anscombe (1967, p. 6).

(v) "The conclusion is that, in the conventional paradigm, the *model* selects the estimator and in turn the *data* selects the estimate. But, in the present case, it is not very realistic to conceive of a model except in a very loose sense. And so one arrives at a tentative notion of what is needed: the *data must select the estimator and with it the estimate*." Nimmo-Smith (1972).

ACKNOWLEDGEMENTS

I am grateful to Dr R. H. Gooding for introducing me to the satellite data, to Mr D. G. King-Hele for allowing me to use it and to Dr Lai Chan, Mr A. P. Dawid and referees for helpful suggestions.

REFERENCES

ALLEN, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–475.

ANDERSON, R. L., ALLEN, D. M. and CADY, F. B. (1972). Selection of predictor variables in linear multiple regression. In *Statistical Papers in Honor of George W. Snedecdor* (T. A. Bancroft, ed.). Ames, Iowa: Iowa State University Press.

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, New Jersey: Princeton University Press.

ANSCOMBE, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *J. R. Statist. Soc.* B, **29**, 1–52.

BARANCHIK, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.*, **1**, 312–321.

BARNETT, V. D. (1966). Order statistics estimators of the location of the Cauchy distribution. *J. Amer. Statist. Ass.*, **61**, 1205–1212.

COCHRAN, W. G. (1968). Commentary on estimation of error rates in discriminant analysis. *Technometrics*, **10**, 204–205.

CURETON, E. E. (1950). Validity, reliability and boloney. *Educ. & Psych. Meas.*, **10**, 94–96.

(1951). Symposium: The need and means of cross-validation. II. Approximate linear restraints and best predictor weights. *Educ. & Psychol. Measurement*, **11**, 12–15.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

EFRON, B. and MORRIS, C. (1973). Combining possibly related estimation problems (with Discussion). *J. R. Statist. Soc.* B, **35**, 379.

GEISSER, S. (1974). A predictive approach to the random effect model. *Biometrika*, **61** (to appear).

GORMAN, J. W. and TOMAN, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27–51.

GRAY, H. L. and SCHUCANY, W. R. (1972). *The Generalized Jackknife Statistic.* New York: Marcel Dekker Inc.

HERZBERG, P. A. (1969). The parameters of cross-validation. Monograph Supplement to *Psychometrika*, **34**.

HILLS, M. (1966). Allocation rules and their error rates (with Discussion). *J. R. Statist. Soc.* B, **28**, 1–31.

HOCKING, R. R. (1972). Criteria for selection of a subset regression: Which one should be used? *Technometrics*, **14**, 967–970.

HOGG, R. V. (1967). Some observations on robust estimation. *J. Amer. Statist. Ass.*, **62**, 1179–1186.

HORST, P. (1941). *Prediction of Personal Adjustment.* New York: Social Science Research Council (Bulletin 48).

JOHNSON, N. L. and LEONE, F. C. (1964). *Statistics and Experimental Design*, Vol. 2. New York, London and Sydney: Wiley.

KATZELL, R. A. (1951). Symposium: The need and means of cross-validation. III. Cross-validation of item analyses. *Educ. & Psychol. Measurement*, **11**, 16–22.

KING-HELE, D. G. and COOK, G. E. (1973). Analysis of 27 satellite orbits to determine odd zonal harmonics in the geopotential. *R. A. E. Technical Report*, No. 73153.

—— (1973). Refining the Earth's pear shape. *Nature*, **246**, 86–88.

KING-HELE, D. G., COOK, G. E. and SCOTT, DIANA W. (1969). Evaluation of odd zonal harmonics in the geopotential, of degree less than 33, from the analysis of 22 satellite orbits. *Planet. Space Sci.*, **17**, 629–664.

LACHENBRUCH, P. and MICKEY, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

LARSON, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.*, **22**, 45–55.

LINDLEY, D. V. (1968). Choice of variables in multiple regression. *J. R. Statist. Soc.* B, **30**, 31–66.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc.* B, **34**, 1–41,

MOSIER, C. I. (1951). Symposium: The need and means of cross-validation. I. Problem and designs of cross-validation. *Educ. & Psychol. Measurement*, **11**, 5–11.

MOSTELLER, F. and TUKEY, J. W. (1968). Data analysis, including statistics. In *Handbook of Social Psychology* (G. Lindzey and E. Aronson, eds), Vol. 2. Reading, Mass.: Addison-Wesley.

MOSTELLER, F. and WALLACE, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist. Ass.*, **58**, 275–309.

—— (1964). *Inference and Disputed Authorship: The Federalist.* Reading, Mass.: Addison-Wesley.

NICHOLSON, G. E. (1960). Prediction in future samples. In *Contributions to Probability and Statistics* (I. Olkin *et al.*, eds). Stanford University Press.

NIMMO-SMITH, IAN (1972). Selecting an estimator from a class of estimator types: an embryo data-analytic technique. M.Sc. dissertation, University of London.

SIMON, FRANCES H. (1971). *Prediction Methods in Criminology.* Home Office Research Study, **7**. London: H.M.S.O.

STEIN, C. (1960). Multiple regression In *Contributions to Probability and Statistics* (I. Olkin *et al.*, eds). Stanford University Press.

—— (1962). Confidence sets for the mean of a multivariate normal distribution. *J. R. Statist. Soc.* B, **24**, 265–296.

TEATHER, D. (1974). Statistical techniques for diagnosis. Submitted to *J. R. Statist. Soc.* A.

WHERRY, R. J. (1931). A new formula for predicting the shrinkage of the multiple correlation co-efficient. *Ann. Math. Statist.*, **2**, 440–457.

—— (1951). Symposium: The need and means of cross-validation. IV. Comparison of cross-validation with statistical inference of betas and multiple *R* from a single sample. *Educ. & Psychol. Measurement*, **11**, 23–28.

DISCUSSION OF PROFESSOR STONE'S PAPER

Professor G. A. BARNARD (University of Essex): I am sure we are grateful to Professor Stone for drawing our attention to a general procedure which is finding increasing-application in data analysis and for formulating a general characterization of this procedure so that we can attempt to place it in proper relation to the more standard methods. The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number of cases where it is actually used.

In spite of its advocacy by distinguished people, such as Deming and Mahalanobis, it remains the case that the vast majority of survey results, for example, are published either without any assessment of sampling errors or with wholly inappropriate estimates based on simple binary calculations. Such practices may have been tolerable when surveys were used in an indicative sense only, to give rough estimates, but nowadays with more and more apparently sophisticated computer programs for social science, failure to take account of possible sampling fluctuations is leading to a glut of unsound analyses which already seems to be bringing statistical method in that area into unjustified disrepute. I have in mind procedures such as AID, the automatic interaction detector, which guarantees to get significance out of any data whatsoever. Methods of this kind require validation, or rather assessment, by some method such as the split sample technique before anyone takes any notice of its results.

I would like to say a few words about the distinction which Professor Stone draws between cross-validation and jack-knifing. In my opinion the term "jack-knifing" should be used for both procedures. As you know, its etymology refers to the jack-of-all-trades and the master of none, and is intended to denote a tool which has a very wide application but which can in most cases be improved upon by more special-purpose equipment.

Another characteristic feature of the procedures is that they appear to be applied in situations where one may have serious doubts of their appropriateness. For example, Lachenbruch and Mickey (1968) applied this cross-validation procedure to the classification of individuals into two classes. My experience of this problem is limited, but I expect many of you would agree with me that such zero-one classifications are almost never what is required in a practical situation. Where it seems to be required, further thought often shows that a better answer will be provided not by a zero or a one, a rigid classification of the individual, but by a continuous variable P which may be interpreted as a probability of belonging to class one. For example if one is trying to estimate the relative frequency of class one, then the average value of P will give you a better estimate than a straightforward count, and in fact better estimates still are usually to hand. In this context, Kerridge's work deserves to be better known. In all these cases, therefore, I think a further analysis of what the problem really is, using some kind of model or, as Kerridge pointed out, a very wide set of possible models, will give you a better answer than a simple straightforward direct attack on the problem as originally stated.

This, I think, is a defect in cross-validation procedures in that they tend to assume the form in which the answer to the problem should be given. Alternatively, they take it for granted and get an answer in that form when it may very well be that an answer ought to be given in a different form. While it is possible to extend cross-validation procedures to allow for various forms of answer I think it would extend the method too far beyond its natural domain to do this. What, you may ask, is its natural domain? It seems to me that an answer is to be found in comparing the two block diagrams of the paradigms of cross-validation on the one hand and the conventional statistics on the other. The important difference in the present paper is in that the block diagram for the conventional paradigm has an outlet at the bottom which leads to further data, whereas the block diagram for cross-validation does not.

It is true that we sometimes have to make do for a very long time with just the data that we have to hand and have very little hope of getting any more. We should always bear in mind that the real basis of statistics, and of science in general, is always the possibility at least, and usually indeed the actuality, of further repeated observations. The repeatable experiment is I think the foundation of physical science, and I hope in time it will become the foundation of social science: and, of course, it lies at the foundation of classical statistics.

Consequently, my feeling about the role that cross-validation procedures should play resembles my view of the role that non-parametric methods should play; they should be resorted to only when we are driven and they must surely be dependent on standard modelling methods for suggestions as to forms of plausible prescriptions. Who, for example, would take seriously the shrinkers that Professor Stone has put before us if it were not for the fact that we all have seen very good reasons for supposing that in many such situations shrinkers are called for. I think the argument presented by cross-validation appears to be very weak in Example 3.1 (p. 117) when one extends a simple model like $\hat{y}(\alpha, S) = \alpha\bar{y}$ to what appears to be an improvement, $\hat{y}(\alpha, S) = \alpha_1 + \alpha_2 \bar{y}$, and then one finds that one loses the shrinker. This sort of behaviour is a bit startling, and it is difficult to understand in any sort of general context unless one refers to the model approach.

I must disagree with Professor Stone's suggestion that something like cross-validatory methods might be used when we are asked to estimate the location of a sample which comes to us without a pedigree. The correct thing to do with the data without a pedigree is to send the mongrels back. If we have decided not to do this and we have some sort of pedigree, then a classical mixed model approach based on such of the pedigree as we can dig out, allowing for a reasonably wide range of possible forms for the sampling distribution methods, and a sketch of the associated likelihood contours (taking account of the actual sample we have to hand), will give us far more information and a better answer than any suggested cross-validation procedure. Incidentally, in some situations when we think we might have an answer in terms of a single number, an approach on these lines may suggest an answer in the form of two alternative locations for the sample.

In closing my remarks I hope that Professor Stone will put into the printed text the further details of the satellite problem which he has given us tonight. Although we are probably more interested in the state of the world than the shape of it, nonetheless I think it is most important that statisticians should be thought of as being interested in the scientific content of the problems they handle. It seems from the diagram we were shown tonight that the South Polar Ice Cap is doing what one would expect it to do. (I assume that the diagram relates to an earth which is a pseudo-earth of uniform density and not the actual physical earth.) Now I would support the approach of King-Hele and Cook rather than cross-validation and it seems to me that the coincidence of analysis is not entirely but to rather a large extent coincidental. Concerning the example on p. 127, surely cross-validation gives us the wrong answer, or at least a rather unlikely answer. We do not know I suppose what the true answer was but the fact that the standard method picks out as significant the three-factor interaction *ADE* when these are just the three factors which give the largest main effects seem to me to be significant and important and I would certainly tend to pay attention to it, whatever Professor Stone would do, whereas cross-validation would not show it to me at all.

To sum up and perhaps pick up Professor Stone's reference to the Prime Minister, that while I think Professor Stone has shown that cross-validation may not have an unacceptable face it still seems to me that it has a shaky fundament.

It gives me pleasure to propose the vote of thanks.

A. C. Atkinson (Imperial College): In this evening's interesting paper Professor Stone has been concerned with methods of data analysis for prediction. Although this seems reasonable in many of his examples, I am not sure that it is appropriate for the analysis of the pilot plant data. In such factorial cases the interpretation of the individual coefficients is often of primary concern.

A strange feature of these data is that not only are the effects *A*, *D* and *E* significant, but so is the three-factor interaction *ADE*. The half-normal plot on p. 189 of Johnson and Leone (1964) shows this pattern clearly. To find a simpler representation of the data I tried a Box and Cox power transformation with an additive model. The approximate $C(\alpha)$ test for this family (Atkinson, 1973) suggests a power less than one and the maximum-likelihood estimate is near one-third. This transformation achieves a slight reduction in the three-factor interaction but does not lead to a satisfactory additive model, a result which suggests the presence of some alternative structure in the data. The elucidation of this structure seems to me an important aspect of any further data analysis.

I am sorry that Professor Stone did not give any examples of the application of his procedures to discrete data. I would like to see an analysis of, for example, binary data or of counts with some structure in the means. Has Professor Stone any comments on or experience of such analyses?

Some robust estimates of location of a random sample are derived in Section 3 of the paper. It would be interesting to know whether cross-validation can be extended to deal with other problems of robust estimation. Since the procedures involve calculating the effect of deleting one observation at a time, it should be possible to extract some information on the presence of outliers.

I have much pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Dr Lai K. Chan (University College London and University of Western Ontario): The application of cross validation in Example 3.1 provides a fresh method for constructing robust estimators. The following is a comparison of the Cross estimator (3.2) determined by $L = |y-\hat{y}|$ with the seven Hampel's *M*-estimators given in Andrews *et al* (1972, Section 2C3). These *M* estimators are considered to be very promising robust estimators of the location parameter. For an estimator having a symmetric distribution, define its pseudo-variance $\equiv n(2{\cdot}5$ per cent point of the distribution$)^2/(1{\cdot}96)^2$. (Pseudo-variance is a more robust measure of dispersion than variance.) When the parent distribution is normal or Cauchy, Cross is always better than the *M*-estimators with $n = 5$. Comparing to the *M*-estimators with $n = 10$, Cross is better than and about the same as 4 and 3 respectively, of these estimators when the distribution is normal and is better and worse than 1 and 6, respectively, of these estimators when the distribution is Cauchy. The inferiority of the last case may be caused by the fact that the prescription (3.2) does not trim off extreme Cauchy observations.

Cross is also similar to the so-called adaptive estimators, i.e. estimators which adapt themselves to the observations of the particular sample available. Comparison of the pseudo-variances of Cross with the eleven adaptive estimators with $n = 5$ in Andrews *et al.* (Sections 2B3 and 2E1) shows that Cross is always better than these estimators when the distribution is Cauchy and is better than and about the same as 8 and 3, respectively, of these estimators when the distribution is normal. It is better than, about the same as and worse than 5, 5 and 1, respectively, of these estimators with $n = 10$ when the distribution is normal; it is better and worse than 9 and 2, respectively, of these estimators with $n = 10$ when the distribution is Cauchy.

To compete with other robust estimators according to the criteria of pseudo-variances or efficiencies, selections of prescription and loss function may have significant effect on the position of Cross. Here are some proposals.

If it is given that the unknown true parent distribution is a member of a family $\mathscr{F}$, say for the purpose of illustration $\mathscr{F} = \{\text{Normal, Cauchy}\}$, we can let

$$\hat{y}(b, S) = b\hat{\theta}_{\mathrm{N}} + (1-b)\,\hat{\theta}_{\mathrm{C}},$$

where $\hat{\theta}_{\mathrm{N}}$ and $\hat{\theta}_{\mathrm{C}}$ are some highly efficient estimators of the Normal and Cauchy location parameters, respectively. (If we are less certain about the tails of the distributions in $\mathscr{F}$, estimators like best linear unbiased estimators based on selected sample quantiles can be used.) Let the defining loss function be $|y-\hat{y}|^{a}$. Then apply Double Cross to choose $\alpha^{\dagger\dagger}(S) = \{a^{\dagger\dagger}(S), b^{\dagger\dagger}(S)\}$. Incidentally, it would be interesting to see if $\hat{y}$ has the following important property of the Birnbaum–Laska–Miké optimally efficient Pitman type estimators (Birnbaum and Miké, 1970): b converges to 1 as n tends to infinity when normal is the true distribution, say.

When $\mathscr{F} = \{F_k\}$ is large and the shapes of the distributions in $\mathscr{F}$ are not all very close, the choice of $b^{\dagger\dagger}$ becomes computationally intractable. One possible solution is to let

$$\hat{y}\{(a, b), S\} = \sum_{k} a_k\, b_k\, \hat{\theta}_k / (\sum a_k\, b_k),$$

where $a_k = 0$ or 1 correspond to the "choice of the variables" in Example 3.4. $\sum_k a_k = 2$ or 3 if each time two or three distributions are to be considered. Alternatively, instead of linear combinations of $\hat{\theta}$'s, we can let b be the characterizing parameter of some widely used robust estimators such as the trimming proportion of trimmed or Winsorized means (so Cross can be considered as a refinement of these estimators).

Like most robust estimators, analytical justification of the above proposals may be difficult when n is finite. But systematical Monte Carlo studies and heuristic deduction can provide useful guidelines.

Another possible interesting application of "cross-validation" is robust estimation of scale parameters. For this, transformation of y may be needed to make "prediction" meaningful.

Mr A. P. Dawid (University College London): Professor Stone has emphasized cross-validatory choice at the expense of cross-validatory assessment, although his analysis clearly brings out the latter as fundamental. The question of choice of predictor is likely to remain a controversial one, but a valid method of assessing the worth of any chosen prediction formula will be welcomed by all statisticians. So the vital question is: How reliable is cross-validatory assessment?

Let us take as given a loss-function $L(y, \hat{y})$. Let $p = \{p_S(\cdot)\}$be some method of prediction, to be assessed, which for each set of observed data S yields a *predictor* $p_S(\cdot)$. Thereafter, if a value for x is observed, the prediction $\hat{y}$ of y is taken to be $p_S(x)$. The loss incurred will, of course, be $L(y, p_S(x))$.

Professor Stone has carefully avoided probability models throughout his paper, but for reasoned judgment of his methods we must make explicit some concepts implicit in his work. I shall assume that the past data S, and the future data (x, y), form a random sample from some unknown distribution, and use a tilde to denote random variation according to that distribution. Then a quantity of interest would be the *predictive risk* $R(p_S) = E[L(\tilde{y}, p_S(\tilde{x}))]$ of the predictor p_S based on data S.

Considering Example 2.1, with squared-error loss, let p correspond to the "naive choice" $\alpha = 1$. Thus $p_S(x) = \bar{y} + (Sxy/Sxx)(x-\bar{x})$, with $Sxy = \sum (x_i - \bar{x}) y_i$, etc. If we then suppose that $(\tilde{x}, \tilde{y})$ have a bivariate normal distribution, with $\tilde{x} \sim N(\mu_x, \sigma_x^2)$ and $\tilde{y} \mid \tilde{x} \sim N(\alpha + \beta x, \sigma^2)$, we find

$$R(p_S) = \sigma^2 + \left\{\left(\alpha - \bar{y} - \frac{Sxy}{Sxx}\bar{x}\right) + \left(\beta - \frac{Sxy}{Sxx}\right)\mu_x\right\}^2 + \left(\beta - \frac{Sxy}{Sxx}\right)^2 \sigma_x^2.$$

Of course $R(p_S)$ depends on the parameters (indeed, on the form) of the unknown distribution. What we require, if possible, is an estimate of $R(p_S)$, as assumption-free as possible.

The "most primitive" form of assessment yields such an estimate. If we take new data $S' = \{(x'_i, y'_i): i = 1, \ldots, n'\}$ from the same distribution as $(\tilde{x}, \tilde{y})$, we can form $L'(S, S') = (1/n') \sum_{i=1}^{n'} L(y'_i, p_S(x'_i))$. Then $L'(S, S')$ will be unbiased and consistent for $R(p_S)$, in the sense that $E\{L'(S, \tilde{S}') \mid S\} = R(p_S)$ and $\operatorname{var}\{L'(S, \tilde{S}') \mid S\} \to 0$ as $n' \to \infty$.

What, though, if we cannot take a new sample, and do not wish to set aside any of our data to yield such an assessment sample? We must construct some method $A = A(p, S)$ of assessment for p. What we should really like is for $A(p, S)$ to be always close to $R(p_S)$—for example, that $E\{\{A(p, \tilde{S}) - R(p_{\tilde{S}})\}^2\}$ should be small; but this seems rather too much to expect. We might have to be satisfied with a good estimator of, not the relevant predictive risk $R(p_S)$, but the *average predictive risk* $r_n(p) = E\{R(p_{\tilde{S}})\}$. (Another formula is $r_n(p) = E\{L(\tilde{y}, p_{\tilde{S}}(\tilde{x}))\}$). Remember $\tilde{S}$ is supposed to arise as a random sample of size n. Returning to our extension of Example 2.1, we find

$$r_n(p) = \frac{(n+1)(n-2)}{n(n-3)}\sigma^2.$$

Professor Stone's cross-validatory assessment method would appear to be a distribution-free method of estimating $r_n(p)$.

Consider firstly the "naive" assessment $\bar{L}(p, S) = (1/n) \sum_{i=1}^{n} L(y_i, p_S(x_i))$. This may be written $\bar{L}(p, S) = E[L(\tilde{y}, p_S(\tilde{x})) \mid (\tilde{x}, \tilde{y}) \in S]$, so that $E(\bar{L}(p, \tilde{S})) = E[L(\tilde{y}, p_{\tilde{S}}(\tilde{x})) \mid (\tilde{x}, \tilde{y}) \in \tilde{S}]$. There is no reason why this should resemble $E[L(\tilde{y}, p_{\tilde{S}}(\tilde{x}))] = r_n(p)$. For instance, Example 2.1 yields $\bar{L}(p, S) = RSS/n$, so that

$$E\{\bar{L}(p, \tilde{S})\} = \left(\frac{n-2}{n}\right)\sigma^2 < \sigma^2 < \frac{(n+1)(n-2)}{n(n-3)}\sigma^2 = r_n(p).$$

For cross-validatory assessment we use $C(p, S) = (1/n) \sum_{i=1}^{n} L(y_i, p_{S-i}(x_i))$, which may be written as $E[L(\tilde{y}, p_{\tilde{S}_0}(\tilde{x})) \mid S; \{\tilde{S}_0, (\tilde{x}, \tilde{y})\} = S]$, where $\tilde{S}_0$, $(\tilde{x}, \tilde{y})$ are supposed to arise independently as random samples of size $(n-1)$ and 1 respectively. Then

$$E\{C(p, \tilde{S})\} = E[L(\tilde{y}, p_{\tilde{S}_0}(\tilde{x}))] = r_{n-1}(p).$$

It follows that this assessment yields an unbiased estimator of $r_{n-1}(p)$, which will hopefully (if the prediction method depends "smoothly" on sample size) give a good idea of $r_n(p)$.

Considering that cross-validation does appear to lead to unbiased estimation, I wonder if the links with jack-knifing may not be stronger than Professor Stone concedes.

The result $E\{C(p, \tilde{S})\} = r_{n-1}(p)$ may be verified by straightforward but tedious calculations for our formulation of Example 2.1. However, if it was agreed that an unbiased estimate of $r_n(p)$ was required, and if the normality assumption could be taken seriously, a better assessment would use $[(n+1)/\{n(n-3)\}]\ RSS$. In any event it may appear more satisfactory to use $[\{(n+1)(n-2)(n-1)(n-4)\}/\{n^2(n-3)^2\}]\ C(p, S)$, which in the normal case estimates $r_n(p)$, rather than $r_{n-1}(p)$. Certainly, if we know that the data-producing distribution lies in a certain class, we may often be able to use this information to improve on the distribution-free assessment given by cross-validation.

Although we do get (approximately) unbiased estimates of $r_n(p)$, I have a worrying feeling that these may not be consistent—that is, $\text{var}\{C(p, \tilde{S}) - r_n(p)\}$ remains bounded below as $n \to \infty$. This property would itself be unimportant if we also had

$$E[\{C(p, \tilde{S}) - R(p_{\tilde{S}})\}^2] \to 0$$

(while var $R(p_{\tilde{S}})$ was bounded below), since $R(p_S)$ is the quantity of real interest—but I find this even harder to believe. If my suspicions have any basis, then one should be wary of over-glib use of cross-validatory assessments, since they may be wide of the mark. Even if we do have consistency, correlations among the $C(p, \tilde{S})$ for different p may cast doubt on the choice of that p, out of a given class, which leads to a minimum for $C(p, S)$.

The framework I have presented is not the only possible one, and indeed does not appear relevant for problems where the x-values arise by design rather than chance. For instance, I cannot see any simple justification for use of cross-validation in the Pilot Plant problem.

Professor F. Downton (University of Birmingham): A current nine-day wonder in the press concerns the exploits of a Mr Uri Geller who appears to be able to bend metal objects without touching them; Professor Stone seems to be attempting to bend statistics without touching them. My attitude to both of these phenomena is one of open-minded scepticism; I do not believe in either of these prestigious activities, on the other hand they both deserve serious scientific examination.

I clearly do not possess the powers of Mr Geller; still less do I understand how to harness the forces he claims to use. I feel the same way about Professor Stone's techniques. My difficulties can most easily be illustrated using Example 3.1 on the location of a single univariate samplẽ. There is something very peculiar about a procedure which gives the same estimator for the "prescriptions" (a): $\hat{y}(\alpha, S) = \alpha$ and (c): $\hat{y}(\alpha, S) = \alpha_1 + \alpha_2 \bar{y}$, but a different one for (b): $\hat{y}(\alpha, S) = \alpha\bar{y}$. Indeed it is the total inconsistency of results from apparently similar "prescriptions", which is difficult to take. For example the "prescription"

$$\hat{y}(\alpha, S) = \alpha_1 + \alpha_2 \sum_{i=1}^{n} g(y_i)/n,$$

where $g(y)$ is any function will lead to the estimator $\bar{y}$ using a quadratic loss function. On the other hand, the "prescription"

$$\hat{y}(\alpha, S) = \alpha_1 + \alpha_2 \sum_{i=1}^{n} g(y_i)$$

leads to quite different results depending on the function $g(y)$. Are there some criteria for choosing a "prescription" which I have missed? Does a "prescription" have to be a plausible estimator in some sense? If so how does one know what it is in a more complex situation?

Again why do we stop at one iteration? Suppose a "prescription" $\hat{y}(\alpha, S)$ leads to the cross-validatory estimator based on $\alpha^+(S)$. Would a linear function of that estimator be a good "prescription" and could we get from that an extra-cross-validatory estimator which would be "better" than that based on $\alpha^+(S)$? The possibilities are endless.

On my present understanding I would summarize the properties of Professor Stone's "prescriptions" as follows:

(i) "Prescriptions", which are logically equivalent from a common-sense point of view, may lead to different estimators when an identical cross-validation procedure is adopted.

Presumably this implies that common sense has little application in the world of "prescriptions" and we shall need to learn a whole new logic of equivalences, where, for example (for a quadratic loss-function and single observation validation),

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} g(y_i) \quad \text{if } \hat{y}(\alpha, S) = \alpha_1 + \alpha_2 \bar{y}, \quad \text{unless } \alpha_1 = 0,$$

and so on.

(ii) The "prescriptions" themselves appear to depend on a great deal of alien folklore (traditional sampling theory appears to be alien to Professor Stone's approach). Why otherwise should we even consider a function of $\bar{y}$ or a linear function of the order statistics as a reasonable "prescription"?

In the present state of the art these are the properties of the "prescriptions" of a witch-doctor not of a doctor: however, in fairness to Professor Stone it should be said that a witch-doctor's "prescriptions" sometimes work if the patient has sufficient faith.

Professor J. DICKEY (University College London): My congratulations to Professor Stone for a paper which promises to be a landmark. I am grateful to him for this opportunity to suggest two separate flow charts, for the two processes of Bayesian Data Analysis and Coherent Personal Inference.

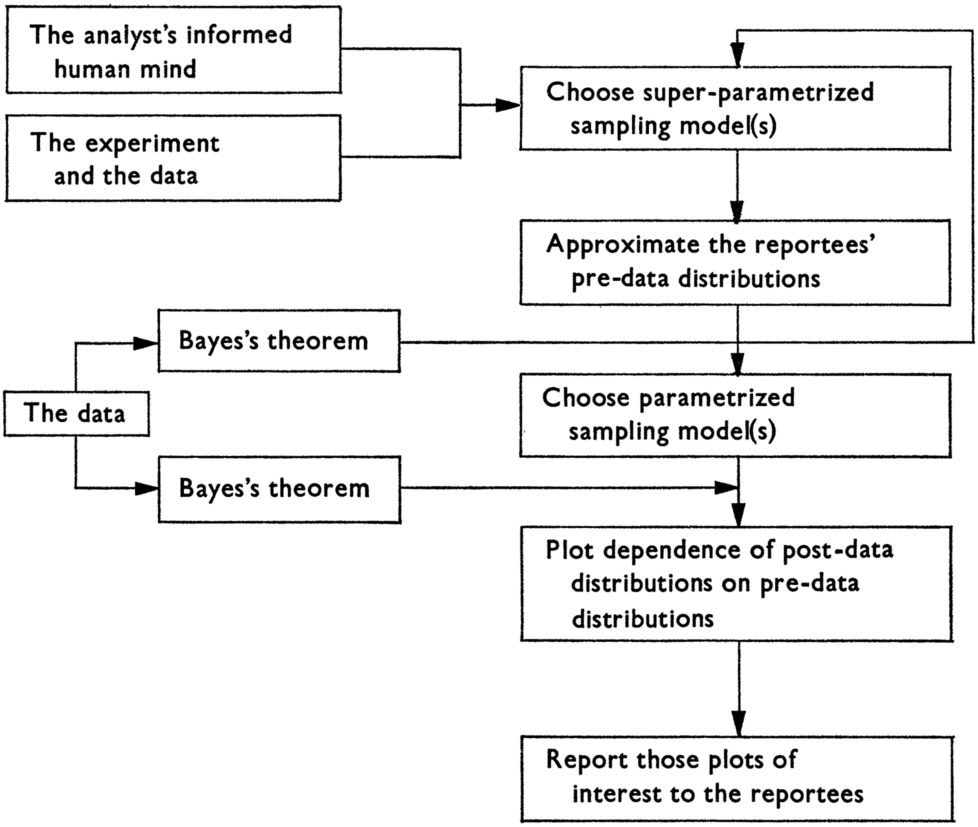
The analyst's informed human mind
The experiment and the data
Choose super-parametrized sampling model(s)
Approximate the reportees' pre-data distributions
Bayes's theorem
The data
Choose parametrized sampling model(s)
Bayes's theorem
Plot dependence of post-data distributions on pre-data distributions
Report those plots of interest to the reportees

FIG. 1. Bayesian analysis and scientific reporting.

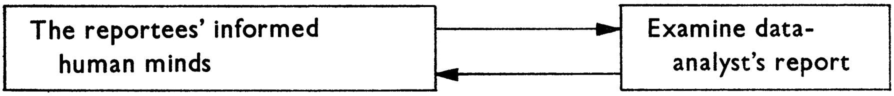
The reportees' informed human minds
Examine data-analyst's report

FIG. 2. Personal inference of reportees.

Mr A. G. BAKER (Unilever Research): As an applied statistician, who seems to become more and more involved in multivariate data, I would like to support Professor Stone in his statement in Section 2: namely that the cross-validatory approach offers the best basis for having a meaningful two-way discussion with a client.

However, the cross-validatory method does have one "unpleasant and unacceptable face". As I see it, the computational costs in a multiple regression problem could become very great. Possibly, it may be best to use conventional methods to cover more problems and only bring in cross-validatory techniques when there is justification for analysing the data exhaustively.

The increased demands on computing would necessitate, in my view, an increasing involvement in the efficiency of computer programs.

My remaining point is not directly linked to cross-validatory techniques, but arises from Professor Stone's example of a 2^5 experiment. Convention has it that we group main effects, first-order interactions, higher order interactions as natural groupings. Another way would be to consider the number of variables involved. The most meaningful equation to describe the results in Table 3 could be

$$\text{response} = f(A, D, E),$$

which gives terms in A, D, E, AD, AE, DE and ADE, i.e. because of the size of ADE: AD, AE and DE would be brought in.

Data description, leading to the formation of hypotheses seems to be the problem in multivariate data and I hope we will get a textbook on cross-validatory techniques coming from University College one day.

I am grateful to Professor Stone for bringing to my notice this work.

Professor O. BARNDORFF-NIELSEN (Department of Theoretical Statistics, Aarhus University): Professor Stone proposes that the quantities C^0, $C^\dagger$ and $C^{\dagger\dagger}$ be employed in the assessment of the respective predictors. But how is the assessment to be carried out? More specifically: Are there methods for grading the scales of value of these quantities, and what are the interpretations of the gradings?

Consider the classical question of predicting y from $y_1, \ldots, y_n$, on the assumption that $y_1, \ldots, y_n$ and y are independent, identically distributed variates following a parametrized family of distributions. It seems desirable that a general scheme like the one put forward by Professor Stone should include the device of estimating the parameter by the maximum-likelihood method and then choosing $\hat{y}$ as that (or those) value(s) of y which has (have) the maximum probability under the distribution with parameter equal to the maximum-likelihood estimate. This is apparently only the case if the loss function at the choice stage of the scheme is allowed to depend on α, a possibility already indicated under point VII of the paper.

Professor D. R. COX (Imperial College): Professor Stone's paper seems to me exceptionally interesting. It represents an approach to statistical analysis that avoids the introduction of parameters, or their equivalent, namely, hypothetical future observations from the same random system. As such it can be contrasted with the recent Scandinavian work in which, roughly speaking, the start is the choice of some statistics (more or less a prescription) and the development centres on the exponential family generated by the statistics.

Two central questions about the present work are (i) to what extent does it give results essentially different from those of more conventional approaches and (ii) how often in applications can the notion of a parameter, or some equivalent, really be avoided?

For example, is the discussion of robust estimation of location essentially different from that of an adaptive estimate in which the sampling variance is estimated for each α and the estimate corresponding to minimum α taken?

Professor Stone's discussion of regression concentrates on the fit of the overall equation and, as such, seems appropriate to the satellite data, but for the factorial experiment interest would normally be more in the nature and interpretation of individual contrasts; can Professor Stone's analysis be easily adapted to this? In the pilot plant example, as Dr Atkinson and Professor Barnard pointed out, the significant contrasts are A, D, E, ADE; also DE is large. This strongly suggests a non-factorial representative of the 8 a–d–e means and, in fact, six of these are about equal and for the other two ad is high and e low. This may provide the most meaningful fit.

Professor SEYMOUR GEISSER (University of Minnesota): When I first learned that Professor Stone and I had at about the same time hit upon the same notion, I was struck by that extraordinary coincidence. After a bit of reflection, it is clear that the idea was inevitable but what is really remarkable, certainly in view of tonight's historical analysis, is that it was so dilatory in coming.

What impelled me to this method was a concern that at the least statistical inference had overemphasized the estimation of non-observables, namely parameters, to the neglect of the prediction of observables or potential observables. I suspect that part of the reason for this is the seductive niceties of the mathematics of parametric structures.

Stress on the estimation of parameters, made fashionable by mathematical statisticians, has been not only a comfortable posture but a secure buttress for the preservation of the high esteem enjoyed by applied statisticians as exposure by observation of any inadequacy in estimation is rendered virtually impossible. On the other hand, predictions can be assessed by further observations or by a sly investigator withholding a part of his data and privately assessing a statistician's predictions. The view that inference should often be couched in terms of prediction of observables, termed predictivism, observablism or, negatively, aparametricism (see Geisser, 1971) is, I am sure, neither new nor held by few. In an even more extreme vein, which possibly only a few share, parameters themselves appear to be for many statistical applications artificial constructs foisted upon an unwary experimenter by self-serving statisticians making "precise" statements about non-existent entities. In fact, I hope that, at the least, Professor Stone's paper makes clear to all that data analytic procedures that flow from his "Cross-validatory" methods (or as I have termed it, stressing the first word, the "Predictive Sample Reuse Method"; Geisser, 1974a, b) are necessary to accommodate those not infrequent instances where both the classical and Bayesian approaches are either too highly structured, too rich or too parameter bound to serve as appropriate vehicles, inferential modes aside.

My version of this low structure predictivistic approach is rather briefly sketched in an applications paper (Geisser, 1974), and set out in greater detail in a recent submission to the *Journal of the American Statistical Association.* It differs from Professor Stone's paper slightly in terminology, although I was pleased to see the addition of the last three words to the title of his paper which were once omitted, giving it now the emphasis in word which was already there in deed. However, I prefer to call the procedure the Predictive Sample Reuse Method because it is essentially a synthesis of the old cross-validatory and curve-fitting approaches with a predictivistic accent. In place of loss function so bound up in decision theory and embodied in *Homo economicus* and "prescription" reminiscent of *Homo medicus*, I favour, on the one hand, the more ambiguous discrepancy function and, on the other, the more direct predictive function, respectively, for *Homo statisticus*. This, of course, may be trivial carping, but I am reminded that a "Sweet William" is also on occasion referred to as a "Stinking Billy" somewhat north of where the present meeting is being held.

A substantive difference in the set-up, as I have considered it, is an allowance for multiple observational omissions. One then may be guided by the particular type of prediction to be made in utilizing a set or a subset of partitions of omitted and retained observations which are relevant to a schema of omissions. In illustrating this, consider Example 3.2 (the k-group problem). Using Professor Stone's prescription and loss function but permitting multiple omissions, formulae for α will vary. For example, if we omit k observations at a time, the total number of possibilities is $\binom{kr}{k}$ of using $kr-k$ retained observations to predict the remaining k, and we would then generate a new estimator for α. If we consider a more natural schema, k simultaneous omissions with exactly one from each group, suggested by simultaneous prediction of a future observation one from each group, the number of possibilities is reduced to r^k and the solution for α is $r/\{(r-1)F+1\}$ if $F\geqslant 1$, and 1 otherwise. This estimator provides greater shrinkage towards the grand mean than the unrestricted one-at-a-time schema of omissions excepting the case when both yield the value 1. The property of greater shrinkage, or its equivalents flattening and centring, with increasing omissions appears to persist in other problems I have recently examined. Another possibility with one-at-a-time omissions is to restrict the omissions to a single group, presuming that prediction for that group only is needed. This again provides a different and interesting answer.

It is clear, of course, that although there is only a finite variety of schemata, including both the quantity of omissions at-a-time and differing patterns of restriction, the number is legion. If we add to this the various possibilities of predictive functions and discrepancy measures we are obviously faced with an *embarras de richesses.* But this should not deter us from making a selection among plausible alternatives guided by the purposes of prediction, cross-validatory assessments and hints from the results yielded by more highly structured modes of inference.

The program of the predictivistic approach to inference requires high structure methods which the Bayesian mode readily yields in terms of predictive distributions of future observations. But data sets, parametric models, prior distributional assumptions being what they entail, the program is incomplete unless adequate low structure predictivistic methods are also available. The cross-validatory or Predictive Sample Reuse Method appears to be a very strong contender for a satisfactory resolution of the low structure situation.

Professor David Hinkley (University of Minnesota): I wish that I could have been present to hear the delivery and discussion of this refreshing paper. I have some questions about the cross-validatory approach, but first let me say how useful I think the approach is.

In Example 3.2, the restriction $0\leqslant\alpha\leqslant 1$ is casually mentioned, presumably since it corresponds to the usual Bayes estimate. For multivariate y's, α is a matrix and the analogous restriction is not easy to handle, cf. Efron and Morris (1972). Should we worry about such parametric restrictions? In the location problem of Example 3.1 it is conventional to take $0\leqslant\alpha\leqslant 1$, but this is often inefficient (in classical parametric terminology) for short-tailed distributions. So here we presumably should drop the conventional restriction.

The central problem in the cross-validation approach is determining an appropriate class of prescriptions. In all of Professor Stone's examples he enlarges the classical prescription to take account of particular types of data structure. But is this too tied to parametric statistics? In the regression problem of Example 3.1 we are using a very special predictor—we should, presumably, always allow for non-linearity, possibly by using an arbitrary bandwidth for a suitable linear smoothing device.

The cross-validatory assessment of prescription relates to samples of size $n-1$ rather than the desired n. Is it often helpful to extrapolate from samples of size $n-2$ and $n-1$

to correct for "overestimation" of loss? Another device that might be considered is the analogue of the Jaeckel epsilon jack-knife, in which observations in S_i each have weight w and (x_i, y_i) has weight $w(1-\varepsilon)$ (rather than zero), the prescription based on this being used to predict the other εw observations with $x = x_i$. Perhaps this ε relates to Professor Stone's mysterious remarks in VII of Section 2 about loss functions.

Professor R. R. Hocking (Mississippi State University): The problem of validation or assessment of a predictor is indeed an important one and it is quite natural to consider partitioning the available data so as to use one portion for estimation (choice of predictor) and the other for assessment of the predictor. The interesting idea proposed in this paper by Professor Stone is that of integrating the procedures of choice and assessment.

I will confine my comments to the problem of selection of predictor variables in linear regression (Example 3.4). In this case the two-stage or double-cross validation leads to precisely the prediction sum of squares (PRESS) proposed by Allen (1971). That is, in terms of the indicator variables a_i,

$$\text{PRESS}\,(\mathbf{a}) = \sum_{i=1}^{n} (y_i - \hat{y}_i(i, \mathbf{a}))^2 = nC^{\dagger}(\mathbf{a}).$$

Here $\hat{y}(i, \mathbf{a})$ is the least-squares estimate of $[Ey_i]$ using a given set of a_i and *excluding* the ith observation. The minimization of PRESS (**a**) with respect to **a** yields a choice of predictor variables from which the usual least-squares predictor is computed.

The choice of predictors selected by double-cross or PRESS may or may not agree with those suggested by other techniques, but in any case it is not uncommon that the resulting predictor fairs poorly on independent observations.

A predictor which appears to do better is the Ridge Regression predictor proposed by Hoerl and Kennard (1970). In this case the predictor is given by $\mathbf{x}'\boldsymbol{\beta}^*$, where

$$\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{X}+\mathbf{D})^{-1}\,\mathbf{X}'\mathbf{Y}.$$

Here **D** is a diagonal matrix with non-negative components a_i on the diagonal. Hoerl and Kennard recommend that $\mathbf{D} = k\,\text{diag}\,(\mathbf{X}'\mathbf{X})$ with the constant k being determined by inspection of the "Ridge Trace", that is, a plot of $\beta_i^*(k)$.

In the present context it is natural to apply the double-cross procedure to the ridge estimator. Thus, with $\mathbf{x}_i'$ denoting the ith row of **X**, the predictor of y_i is $y_i^*(i, a) = \mathbf{x}_i'\boldsymbol{\beta}^*(i, \mathbf{a})$ where for given a_i,

$$\boldsymbol{\beta}(i, \mathbf{a}) = \Big(\sum_{j\neq i} \mathbf{x}_j\,\mathbf{x}_j' + \mathbf{D}\Big)^{-1} \Big(\sum_{i\neq j} \mathbf{x}_j\, y_j\Big).$$

We now determine **a** so as to minimize

$$\text{PRESS}\,(\mathbf{a}) = nC^{\dagger}(\mathbf{a}) = \sum_{i=1}^{n} \{y_i - y_i^*(i, \mathbf{a})\}^2.$$

Having determined **a** we then compute $\boldsymbol{\beta}^*$ as above.

This procedure for determining **a** has been recommended by Allen (1974) who proposes an algorithm for calculating the a_i. With respect to selection of variables, large values of a_i suggest deletion of corresponding variables.

In summary, I would say that while cross-validation only simulates validation, it is better than ignoring it altogether. The concept of integrating choice and assessment is appealing especially if we do not constrain ourselves to classical least squares but, rather, allow extensions such as that illustrated here.

Mr A. S. Young (University College London): The prescription $\hat{y}(x;\, \alpha)$ given in (3.12) is a special form of a linear predictor $\hat{y}(x)$ given by

$$\hat{y}(x) = \sum_{k=1}^{p} \theta_k\, C_k(x),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}} \in R^p$. When A is hierarchical, (3.12) is the result of assuming that there is one, and only one, r with $r = 1, \ldots, p$, for which $\boldsymbol{\theta} \in \Psi_r$ where the Ψ_r's are p mutually exclusive p-dimensional subsets of R^p given by

$$\Psi_r = \{\boldsymbol{\psi}_r = (\psi_1, \ldots, \psi_r, 0, \ldots, 0)^{\mathrm{T}} | \psi_1, \ldots, \psi_r \text{ are different from zero}\}.$$

This assumption assigns a special status to $\theta_k = 0$ as apart from other values of θ_k. The consequence is that the estimation of $\boldsymbol{\theta}$ can proceed in two stages: the true r (or **a** in the paper) is determined and then $\psi_1, \ldots, \psi_r$ (or **b**) are estimated. Least-squares estimation, and hence prediction, is considerably improved by this two-stage process.

A reasonable belief about the θ_k's in the hierarchical model is that they will tend to decrease in absolute value as k increases. We express this opinion in distributional form by assuming that for given τ^2 and λ^2

$$\theta_j \sim N[0, \tau^2/\lambda^{2(j-1)}\{(j-1)!\}^2], \quad j = 2, \ldots, p,$$

and they are all independent. We assign a vague prior to θ_1—the constant term—independent of the other θ's and give suitable priors to τ^2 and λ^2. $C_k(x)$ is assumed to be the normalized Legendre polynomial of order $(k-1)$ with x in the interval $[-1, 1]$. We suppose that we have a normal regression model with constant variance σ^2 to which we give a suitable prior.

The prior distribution of $\theta_2, \ldots, \theta_p$ can be deduced from giving the scaled coefficients $\lambda^{(j-1)}(j-1)!\, \theta_j$ an exchangeable (in this case, independent) distribution with a $N(0, \tau^2)$ marginal density. Since as j increases, $\lambda^{(j-1)}(j-1)!$ becomes an increasing function of j, the tendency of $|\theta_j|$ to decrease is thus given a probabilistic expression. The form $\lambda^{(j-1)}(j-1)!$ is chosen to ensure that if $\hat{y}(x)$ is expressed in powers of x, the coefficients of these powers will exhibit this tendency to decrease in absolute value also. We exclude the constant term θ_1 from this decreasing relationship for obvious reasons.

We get a joint posterior distribution of $\boldsymbol{\theta}$ and λ^2 from which we obtain point estimates of $\boldsymbol{\theta}$ and λ^2 by solving the modal equations iteratively. These equations have very simple forms.

The optimal way to predict using these estimates has been determined by Lindley (1968). Beale's comment, in the discussion to that paper that "all the regression coefficients are non-zero" when normal priors are used is not too important. For, depending on the prior given to λ^2, our normal prior could lead to estimates very close to zero for most of the high-order coefficients. These estimates would be so tiny that the practical effect on the value of $\hat{y}(x)$ would be nil for values of x within our range of interest. This eliminates the problem of fitting a $(n-1)$st degree polynomial to n points since the estimated high-order coefficients of this polynomial would be effectively zero.

Whether the assignment of special status to $\theta_k = 0$ would give improved Bayesian estimates in general is debatable; what we have shown is that it is not essential in a hierarchical model. Finally, Halpern (1973) has given a Bayesian predictor for this model similar to Professor Stone's model-mix prescription where the α's are either prior or posterior weights attached to the m models.

The author replied in writing as follows:

Professor Barnard proposes to smudge the distinction I have drawn between cross-validation and jack-knifing. Admittedly some distinctions are worthless, such as that between integrated likelihood and Bayesian methods, but Professor Barnard's novel single-aspect principle of nomenclature would have us all Adams and Eves, even Adeves if we gave up *sex* for *humanity*. Perhaps we could compromise with *jackboot* or even *jockstrap* if some smudging has to be done.

The "future data" box was omitted from Fig. 1 to emphasize that its presence there is less essential than it is in Fig. 4 and not to denigrate its obvious retrospective value.

The prescription (c) of Example 3.1 does not shrink because, unlike (b), it gives special status to no point on the axis.

For an example of application to discrete (binary) data, I refer Dr Atkinson to the baseball data analysis of Example 3.2. I have since applied the model-mix method to multinomial frequencies with a possible structure of "equality of probabilities"; the outcome is a smoothing formula that comes close to that of Good (1965, p. 33). Dr Atkinson's suggestion that the method might have something to contribute to the treatment of outliers is an excellent one.

I am glad that Professor Barnard, Dr Atkinson, Professor Cox and Messrs Dawid and Baker have raised the question of the proper approach to the Pilot Plant data since this gives me the opportunity to emphasize that the cross-validatory method should not set itself up as a theory of inference but should be content to act as substitute for some of the predictive applications of inference. All that our analysis of Table 3 has told us is that, if we wish to restrict ourselves to one of the four predictors listed, the indications are that we should ignore the lonestar of the *ADE* interaction. I suggest that this information is useful whether or not we do so restrict ourselves. Can Dr Atkinson and Professor Cox offer any assessment, based on the data, of the predictive value of their approaches? A prescription that is suggested by the plausible criticisms of the prescription I first thought of is:

$\hat{y}(x;\, \alpha, S)$ = least-squares predictor for a linear model for factors with interactions possibly up to but not above the three-factor interaction level

$\mathscr{A}$ = {the subsets of A, B, C, D, E}

L quadratic

Using this we obtain $\alpha^{\dagger}(S) = \{A, D, E\}$ and $C(\alpha^{\dagger}(S)) = 5{\cdot}1$, which is not much less than the minimum value, 5·9, of $C^{\dagger}(\mathbf{a})$ for the first prescription. (I have followed the hint at the end of Example 3.5 and have dispensed with "double-cross" by incorporating least-square predictors directly into the prescription.) Nevertheless, a smaller minimum is to be expected *a priori* since the second prescription, having 32 elements in its $\mathscr{A}$, is probably more adaptive and possibly more realistic. For both prescriptions, it would be interesting to have the value of the cross-validatory *assessment* measure. (The analysis of Example 3.6 was incomplete in this respect because a computer program for the calculation of this measure has not yet been written.) It would be "very interesting" if, as seems quite possible, we obtained assessments favourable to our first uninspiring prescription and unfavourable to the second, apparently more acceptable, one. The second prescription corresponds to Professor Cox's analysis, unless, that is, he wishes to do something about his observation of near equality of 6 of the *a–d–e* means; for the cross-validation criterion is the same whether we think of the eight means as having a factorial structure or not. The present approach may help in the interpretation of the meaning of Cox's phrase "the most meaningful fit".

I am grateful to Professor Chan for taking up the comparison question at the end of Example 3.1 and using his expert knowledge of the area to show that Cross should, at least, be encouraged to run for its country in the Robustness Open.

Cross-validation will undoubtedly benefit from *some* theoretical underpinning. Few can be better equipped to attempt this than Mr Dawid. His first results are interesting and may go some way to answer Professor Barndorff-Nielsen's pertinent questions.

Professor Downton thinks that his examples place cross-validation with the black arts. My interpretation of them is precisely the opposite. That cross-validatory choice, applied to the prescription $\alpha_1 + \alpha_2 \sum g(y_i)/n$, $\mathscr{A} = R^2$ and L quadratic, yields $\hat{y} = \bar{y}$ is very reassuring, when perverse possibilities such as $g(y) = y^{100}$ are borne in mind. Try to do something silly and the method will do its best to stop you, will even succeed in doing so.

Far from being black art, this is more guardian angel. The fact that the method fails, indeed runs amok, with the presdription $\alpha_1+\alpha_2 \sum_{i=1}^n g(y_i)$ is surely no cause for censure, since the individual predictors, prior to choice of α, have a dependence on sample size that would rule them out on any statistical principles.

Professor Dickey has struck a blow against any simplistic account of the Bayesian position. I accept his implied rebuke and will stop making any predictions about the viewpoint of Bayesians.

Mr Baker's suggestion that, where computational costs loom large, cross-validatory calculations may be reserved for exhaustive analyses only is, at first sight, attractive. However, I do not think that the essential benefits of the method can be realized if it is regarded as an optional extra to the conventional methods. My experience has been that, for multiple regression, the C.P.U. time of the method is not very much greater than that of the standard methods, even with my fortranical blunderings. Moreover C.P.U. costs are falling while output costs are increasing; by cutting down on "side-inch", the method may come to merit an ecological halo.

Incidentally, I must disclaim any responsibility for the epithet "best" in Mr Baker's encouraging opening, recalling Voltaire's dictum that *mieux* and *bien* make poor bedfellows, a dictum whose cogency of application to Statistics is underlined by Tukey's resistance to "the tyranny of the best".

Being not enamoured of extraordinary coincidences, I am inclined to assume that I am the honoured target of subliminal statistical stimulation from the land of Minnehaha. Indeed, I fancy I am beginning to feel the first rumblings of the Exponential Horn (or Master's Voice) syndrome. Having said which, it seems hardly necessary to add that I warmly agree with most of what Professor Geisser says. For my swan-song of independent thought, however, I conjecture that Geisser is on the wrong track in leaving out more than one item at a time, that whatever diluted optimality theorems exist in this area will require the $n-1$ to 1 split.

Professor Hinkley has raised some very suggestive questions. One can, of course, generalize Example 3.2 without introducing matrix α but the latter possibility is intriguing.

I am grateful to Professor Hocking for the references to some recent work of Allen. My allusion to the challenge of the possibility $p = \infty$ at the beginning of Example 3.4 was to some preliminary skirmishing with a cross-validation approach to the choice element of ridge regression. I am glad to find that this competitor for "model-mix" has been already explored by the "Technometrics school", if I may call them so, and I look forward to reading the findings.

Mr Young's contribution lies in the same area but has a less empirical basis. I cannot yet decide whether the coincidence of some Bayesian and cross-validatory results is complimentary to the former or the latter. Perhaps "complementary" would be a better word. I hope that Young will eventually obtain the numerical output of the Bayesian approach and, even, dare I suggest it, calculate a cross-validatory assessment for it.

References in the Discussion

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting variables. Technical Report No. 23, Dept. of Statistics, University of Kentucky.

—— (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Atkinson, A. C. (1973). Testing transformations to normality. *J. R. Statist. Soc.* B, **35**, 473–479.

Birnbaum, A. and Mike, V. (1970). Asymptotically robust estimators of location. *J. Amer. Statist. Ass.*, **65**, 1265–1282.

Efron, B. and Morris, C. (1972). Empirical Bayes and vector observation: an extension of Stein's method. *Biometrika*, **59**, 335–347.

Geisser, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 456–469. Toronto, Montreal: Holt, Rinehart & Winston.

GOOD, I. J. (1965). *The Estimation of Probabilities.* Cambridge, Mass.: M.I.T. Press. (Research Monograph No. 30.)

HALPERN, E. F. (1973). Polynomial regression from a Bayesian approach. *J. Amer. Statist. Ass.*, **68**, 137–142.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–68.

Added by the author in proof:

FISHER, R. A. (1924) III. The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. Roy. Soc. London* B, **213**, 89–142.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores.* Chapter 13. Reading, Mass.: Addison–Wesley.