J. R. Statist. Soc. A (1984),
147, *Part* 3, *pp.* 389–425

Selection of Subsets of Regression Variables

By ALAN J. MILLER

CSIRO Division of Mathematics and Statistics,
Melbourne, Australia

[*Read before the* Royal Statistical Society on *Wednesday, January 25th, 1984, the President*, Professor P. Armitage, *in the Chair*]

SUMMARY

Computational algorithms for selecting subsets of regression variables are discussed. Only linear models and the least-squares criterion are considered. The use of planar-rotation algorithms, instead of Gauss–Jordan methods, is advocated. The advantages and disadvantages of a number of "cheap" search methods are described for use when it is not feasible to carry out an exhaustive search for the best-fitting subsets.

Hypothesis testing for three purposes is considered, namely (i) testing for zero regression coefficients for remaining variables, (ii) comparing subsets and (iii) testing for any predictive value in a selected subset. Three small data sets are used to illustrate these tests. Spjøtvoll's (1972a) test is discussed in detail, though an extension to this test appears desirable.

Estimation problems have largely been overlooked in the past. Three types of bias are identified, namely that due to the omission of variables, that due to competition for selection and that due to the stopping rule. The emphasis here is on competition bias, which can be of the order of two or more standard errors when coefficients are estimated from the same data as were used to select the subset. Five possible ways of handling this bias are listed. This is the area most urgently requiring further research.

Mean squared errors of prediction and stopping rules are briefly discussed. Competition bias invalidates the use of existing stopping rules as they are commonly applied to try to produce optimal prediction equations.

Keywords: SUBSET SELECTION; MULTIPLE REGRESSION; VARIABLE SELECTION; MEAN SQUARED ERRORS OF PREDICTION; MALLOWS' C_p; AKAIKE'S INFORMATION CRITERIA; PREDICTION; LEAST SQUARES; CONDITIONAL LIKELIHOOD; STEPWISE REGRESSION

1. INTRODUCTION, OBJECTIVES, STRATEGIES

The extensive literature on selecting subsets of regressor variables was well reviewed by Hocking (1976). The literature, and Hocking's review, are largely on (i) computational methods for finding best-fitting subsets, usually in the least-squares sense, and (ii) mean squared errors of prediction (MSEP) and stopping rules. Hocking also discusses alternatives to using subset methods, such as using ridge regression, and using subsets of orthogonal linear combinations of all of the available predictor variables. There is little on inference or estimation, though in several places Hocking mentions that standard least-squares theory is not applicable when the model has not been determined *a priori*, e.g.

> "The properties described here are dependent on the assumption that the subset of variables under consideration has been selected without reference to the data. Since this is contrary to normal practice; the results should be used with caution."

Present address: Private Bag 10, Clayton, Victoria, Australia 3168.

0035–9238/84/147389 $2.00

Thompson (1978) has also reviewed subset selection in regression, and Hocking (1983) has reviewed developments in regression as a whole over the period 1959-82.

As stepwise regression is one of the most widely used of all statistical techniques, an examination of its methods and "folklore" is long overdue. Most of the criticism of subset regression methods in this paper applies also to the fitting of time series models in which the model has not been decided *a priori*. Though only linear models will be discussed here, the same ideas apply to all models, linear or non-linear, where the subset of predictor variables has not been decided *a priori*. Reasons for using only some of the available or possible predictor variables include:

(i) to estimate or predict at lower cost by reducing the number of variables on which data are collected,

(ii) to predict accurately by eliminating uninformative variables,

(iii) to describe a multivariate data set parsimoniously.

(iv) to estimate regression coefficients with small standard errors (particularly when some of the predictors are highly correlated).

These objectives are of course not completely compatible. Prediction is probably the most common objective, and here the range of values of the predictor variables for which predictions will be required is important. The subset of variables giving the best predictions in some sense, averaged over the region covered by the calibration data, may be very inferior to other subsets for extrapolation beyond this region. For prediction purposes, the regression coefficients are not the primary objective, and poorly estimated coefficients can sometimes yield acceptable predictions. On the other hand, if process control is the objective then it is of vital importance to know accurately how much change can be expected when one of the predictors changes or is changed.

Alternatives to using subset selection which may achieve some of the objectives include ridge regression and other shrinkage methods, the use of subsets of orthogonal (or other) linear combinations of the predictors, factor analysis, etc. Only when the statistical properties of subset regression methods are understood can we hope to compare these alternatives objectively. It may be appropriate to use some kind of shrinkage estimator in conjunction with subset regression.

The many strategies of subset regression can be categorized conveniently by breaking them into the following phrases:

(i) Decide the variable(s) to be predicted and the set of possible predictors, and then assemble or collect a data set.

(ii) Find subsets of variables which fit the data well.

(iii) Apply a stopping rule to decide how many predictors to use.

(iv) Estimate regression coefficients.

(v) Test how well the model fits, examine residuals, etc., possibly adding new variables at this stage (e.g. polynomial terms, interactions, transformations), and returning to phase (ii) above.

The paper by Cox and Snell (1974) provides some useful advice on topic (i) in the context of medical statistics. The most widely used algorithm, due to Efroymson (1960) and often just termed stepwise regression, combines together phases (ii) and (iii) above, using false F-tests as the stopping rule. Computational algorithms are discussed in Section 2 of this paper.

Hypothesis tests, particularly using the F-to-enter statistic, are often used as stopping rules, whether the objective is parsimonious model building or prediction. Inference and stopping rules in prediction are treated separately in Sections 3 and 5 respectively.

The most important unresolved problem is that of estimation. Sources of bias and their treatment are discussed in Section 4.

It will be assumed that we have an $n \times (k + 1)$ matrix X consisting of the n values of k predictor variables, and a column of 1's if a constant is being fitted, and a corresponding $n \times 1$ vector Y of observed values of the variable to be predicted. It will be assumed that the relationship between Y and the predictors is

$$Y = X\beta + e(X) + \epsilon,$$

where β is a $(k + 1) \times 1$ vector of unknown regression coefficients (some of which may have zero value), $e(X)$ is the deterministic error in the linear model and is defined to be orthogonal to X, and ϵ is a vector of true but unknown residuals whose elements have zero expected value and are independently and identically distributed.

2. FINDING SUBSETS WHICH FIT WELL

Algorithms for finding best-fitting subsets of variables to a set of data requires (i) a search strategy, and (ii) a computational algorithm. Fitting linear models by least squares is the simplest case and the emphasis in this paper will be on this method. For this case, a wide range of combinations of search strategy and computational algorithm is available, and the feasibility of various combinations depends upon the numbers of predictors and observations. If non-linear models are being fitted, or the measure of goodness-of-fit is not a sum of squares, the same search strategies are theoretically still available but will often not be feasible because of the greatly increased computational complexity. Some examples of the use of criteria other than least squares include the fitting of log-linear models to categorical data (e.g. Goodman 1971; Brown, 1976; Benedetti and Brown, 1978), minimax or L_∞-fitting (e.g. Gentle and Kennedy, 1978) and minimizing the sum of absolute deviations or L_1-fitting (e.g. Roodman, 1974; Gentle and Hanson, 1977; Narula and Wellington, 1979; Wellington and Narula, 1981).

Search strategies can be divided conveniently into (i) those which guarantee to find the best-fitting subsets of some or of all sizes, and (ii) the "cheap" methods which sometimes find the best-fitting subsets.

Garside (1971a, b) and others have given methods for generating the residual sums of squares for all subsets for all sizes. All of the published algorithms known to the author use Gauss–Jordan methods operating upon sums of squares and products matrices. Alternatively, the planar-rotation algorithm of Gentleman (1973, 1974) can be used to change the order of variables within a triangular factorization, as described for instance by Elden (1972), Hammarling (1974), Dongarra *et al.* (1979) and Clarke (1981). In the author's experience this is only slightly slower (about 25 per cent) than the Garside algorithm but far more accurate, and it can be used in single precision except for the most extremely ill-conditioned data sets. This method cannot return a negative calculated residual sum of squares, and can be used when the number of variables exceeds the number of observations. It is only feasible at present to evaluate the residual sums of squares for all possible subsets for k up to about 20.

There are combinatorial algorithms for generating all subsets of p variables out of k, or all subsets of *p or less* out of k. One such algorithm has been given by Kudo and Tarumi (1974), though the basic algorithms for generating orderings to minimize the computational effort in going from one subset in the sequence to the next can be found in many texts on combinatorial methods (e.g. Reingold *et al.*, 1977; Nijenhuis and Wilf, 1978). If a user has say 50 available predictors, the evaluation of all $(2^{50} - 1) = 10^{15}$ residual sums of squares is not feasible. However in such cases the user may be looking for subsets of say 5 predictors and ${}^{50}C_5 = 2 \times 10^6$, which although a large number of subsets to consider, is just feasible with a fast computer.

If only the best-fitting subset of p predictors is being sought. then many subsets can be skipped using a branch-and-bound technique. This seems to have been proposed first by Beale *et al.* (1967) and by Hocking and Leslie (1967).

The branch-and-bound technique is particularly valuable in reducing the number of subsets to be considered in cases where there are "dominant" predictors such that there are no subsets which fit well which do not contain them. The technique is less useful when the number of predictors exceeds the number of observations. Furnival and Wilson (1974) have described a branch-and-bound algorithm for finding best-fitting subsets of all sizes, using Gauss–Jordan type methods and sums of squares and products matrices. To avoid the substantial computational cost of inverting most of the rows of a matrix to obtain the lower bounds required, they maintain two copies of the matrices, one with only a few rows inverted and the other with the remaining rows inverted.

In searching for say the 10 or 20 best-fitting subsets of each size, then it is the current 10th or

20th best which is compared with the lower bound in deciding whether to eliminate a sub-branch. It is usually feasible to find say the ten best-fitting subsets of p or fewer variables out of k when the total number of different subsets to be considered is up to about 10^8. It is usually an advantage to use one of the "cheap" methods to be described next to find some fairly good bounds so that some of the unprofitable sub-branches can be eliminated early in the search.

The common "cheap" methods such as forward selection, the Efroymson (1960) forward stepwise method, and backward elimination are described in most texts on regression and are well enough known not to merit further description here. Forward selection and the Efroymson algorithm can be used when there are more predictors than observations; backward elimination is usually not feasible in such cases.

There are many who prefer subjective to automatic methods of variable selection. They may perhaps start by finding the simple correlations between the predictand and each of the predictors, and then look at scatter diagrams for those predictors with the largest correlations. This may show the need for a transformation, or adding polynomial terms, or the presence of outliers. After selecting one predictor, the process is repeated using the residuals from fitting this predictor, continuing until nothing more can be seen in the data. This approach is an extension of forward selection and suffers from the weaknesses of that method, though it does provide some protection against the selection of what might be considered stupid models. A formalized version of this procedure has been called "projection pursuit" by Friedman and Stuetzle (1981). Without enumeration of the family of potential models in advance, it is impossible to develop any statistical inference for such methods. The plotting and examination of residuals should of course be a part of any procedure.

An alternative to the Efroymson algorithm, which often finds better-fitting subsets, is that of replacing predictors rather than deleting them. Suppose that we have 26 potential predictors denoted by the letters A to Z and that we are currently looking for subsets of four predictors. Let us start with the subset $ABCD$. Consider first replacing predictor A with that one from the remaining 22 which gives the smallest residual sum of squares in a subset with B, C and D. If no reduction can be obtained then A is not replaced. Then try replacing B, then C, then D, and then back to the new first predictor, continuing until no further reduction can be found. This procedure must converge, and usually converges rapidly, as the residual sum of squares decreases each time that a predictor is replaced, there is only a finite number of subsets of four predictors, and the residual sum of squares is bounded below.

Many variations on the basic replacement algorithm are possible. As described above, the algorithm could converge upon a different final subset if started from subset $DBAC$ instead of $ABCD$, that is if we carry out the replacement in a different order. A variation on the method is to find the best replacement for A, but not to make the replacement. Similarly the best replacements for B, C and D are found but only the best of the four replacements is implemented. The process is repeated until no further improvement can be found. A sequential replacement algorithm is possible, that is it is carried out sequentially for one, two, three, four predictors, etc., taking the final subset of $(p-1)$ predictors plus one other predictor as the starting point for finding a subset of p predictors.

Another variation which is particularly useful when there is a large number of predictors is to use randomly chosen starting subsets of each size. On one problem in the use of near infra-red spectroscopy, there were 757 available predictors of which 6 were to be selected. From 100 different random starts, the replacement algorithm converged upon 74 different final subsets. None of these was the best-fitting subset; an *ad hoc* procedure found one subset of six predictors which gave a residual sum of squares which was only two-thirds of the best found from the 100 random starts.

It should be emphasized that none of these "cheap" procedures guarantees to find the best-fitting subsets. Berk (1978a) constructed an artificial example with four predictors in which forward selection and backward elimination select the same subsets of all sizes, missing a subset of two predictors which gives a residual sum of squares equal to 1/90th of that for the selected subset of two variables. In Berk's example, a replacement algorithm would have found the best-fitting subsets of all sizes.

As an illustration of a case in which forward selection performs badly, consider the artificial data in Table 1. The Y-variable is exactly equal to $(X_1 - X_2)$, but Y is orthogonal to X_1 and almost orthogonal to X_2. Forward selection picks variable X_3 first. A negligible reduction in residual sum of squares is obtained if X_2 is added to X_3 and no reduction occurs if X_1 is added to X_3. Many automatic routines would then stop with only one subset, that containing only X_3 being selected. A similar situation to this often occurs in the physical sciences when a difference between two variables is a proxy for a rate of change in position or time. If such a situation is anticipated, it may be feasible to add all the pairs of differences which seem sensible, to the set of predictors. It will often be desirable though to leave the original variables in the set of available predictors. Some software using poor numerical methods will give problems though if X_1, X_2 and $X_1 - X_2$ are all included.

TABLE 1.
An artificial data set

Observation number	*Predictors*			
	X_1	X_2	X_3	Y
1	1000	1002	0	−2
2	−1000	−999	−1	−1
3	−1000	−1001	1	1
4	1000	998	0	2

All of the "cheap" methods discussed so far have involved adding, deleting or replacing one variable at a time. Algorithms which replace two variables at a time are much slower but they are usually still feasible even with hundreds of predictor variables, and they are more likely to find the best-fitting subsets than one-at-a-time algorithms.

Gabriel and Pun (1979) have suggested that when there are too many predictors for an exhaustive search for best-fitting subsets to be feasible, it may be possible to break the predictors into groups within which the exhaustive search is feasible. The grouping is such that if two predictors A and B are in different groups, then their regression sums of squares are additive. To do this we need to find when regression sums of squares are additive. Is it only when the variables are orthogonal? It can be shown that if r_{AB}, r_{AY}, r_{BY} are the sample product-moment co-relations between the variables A, B and the variable to be predicted, Y, then the condition is that

$$1 - r_{AB}^2 = (1 - r_{AB} r_{BY}/r_{AY})^2 = (1 - r_{AB} r_{AY}/r_{BY})^2 \tag{2.1}$$

which has the two practical solutions $r_{AB} = 0$, i.e. that A and B are orthogonal, and

$$r_{AB} = 2 r_{AY} r_{BY}/(r_{AY}^2 + r_{BY}^2) \tag{2.2}$$

plus the trivial solution $r_{AB} = 1$. At the moment, this is an interesting idea but it has not been implemented.

In many practical cases, in using the above algorithms, some variables will be forced into all subsets, or conditions will be imposed, e.g. that interactions not be included without main effects, etc. Some users also reject subsets if the regression coefficients have the "wrong" signs.

3. HYPOTHESIS TESTING

The three types of hypothesis test which are needed are:

(i) given a subset of predictors (not chosen *a priori*), are the data consistent with zero regression coefficients for all remaining predictors,

(ii) is one subset significantly better than another (e.g. for prediction), and
(iii) is there any predictive value in the selected subset?

Three sets of data will be used for illustration; these are the steam pressure data (STEAM) from Draper and Smith (1981, p. 616), the Detroit homicide data (DETROIT) of Fisher (1976) from Gunst and Mason (1980, p. 360) and the aircraft cost data (PLANES) given by Copas (1983).

In the case of the DETROIT data, the homicide rate per 100 000 population will be used as the variable to be predicted; the accidental death rate and rate of assault will not be used. The predictors for this data set will be numbered from 1 to 11 in the order given in Gunst and Mason starting with the number of police per 100 000 population.

The numbers of predictor variables and the numbers of observations for the three sets of data are:

Data set	*No. of predictors*	*No. of observations*
	(k)	(n)
STEAM	9	25
DETROIT	11	13
PLANES	14	31

The DETROIT data set is remarkable in that the first variable selected in forward selection is the first one eliminated in backward elimination. It is also a case in which a best-fitting subset fits very much better than the subsets of the same size found by forward selection and backward elimination. The best-fitting subset of three predictors gives a residual sum of squares of 6.77 compared with residual sums of squares of 21.2 and 23.5 respectively for the subsets found by forward selection and backward elimination.

A number of methods are in popular use for testing the hypothesis that the regression coefficients are zero for all of the variables which have not been selected. These include using the F-to-enter statistic, the use of added dummy variables, a permutation test of Forsythe *et al.* and using the lack-of-fit statistic. Except for the method using dummy variables, they all require the assumption that the p variables which have been selected at the time that the test is applied have all been chosen *a priori*; the effect of the selection process on these tests does not seem to have been considered. A test due to Spjøtvoll (1972a) does not suffer from this problem and can be used instead of these tests when the number of observations exceeds the number of predictors.

If RSS_p and RSS_{p+1} are the residual sums of squares when linear models in p and $(p+1)$ predictors respectively have been fitted together with a constant, then we can define a variance ratio:

$$VR = \frac{\mathrm{RSS}_p - \mathrm{RSS}_{p+1}}{\mathrm{RSS}_{p+1}/(n-p-2)} \tag{3.1}$$

This is the F-to-enter statistic. In forward selection and the Efroymson algorithm, the $(p+1)$th predictor is that which maximizes VR, so that the F-to-enter is the first order-statistic from a sample of $(k-p)$ variance ratios, or $(n-p)$ if the number of predictors exceeds the number of observations, but where the variance ratios are usually not independent. With other subset selection strategies, the quantity (3.1) may be useful where some or all of the p predictors are not in the subset of $(p+1)$ predictors.

The distribution of the maximum F-to-enter is of course not even remotely like an F-distribution. This was pointed out by Draper *et al.* (1971) and by Pope and Webster (1972). The true distribution of the maximum F-to-enter is a function of the values of the predictor variables. Draper *et al.* (1979) derive the distribution for the case of two orthogonal predictor variables when the wrong one is chosen, with the true value of the regression coefficient being zero for the chosen predictor.

An approximation to the distribution of the maximum F-to-enter can be obtained using an order-statistic argument. If we assume that we have $(k-p)$ independent variance ratios, then the probability, α, that the largest exceeds some value F_0 is

$$\alpha = 1 - (1 - \alpha^*)^{k-p}, \tag{3.2}$$

where α^* is the tail probability for a value of F exceeding F_0; the number of degrees of freedom of F are 1 and $(n-p-2)$. The approximation (3.2) is often used in meteorological applications to provide a stopping rule for subset selection following the recommendation by Miller (1962). If a value of $\alpha = 5$ per cent is chosen, this may be found to correspond to α^* = say 0.1 per cent and hence to using a limiting value for the F-to-enter in excess of 10.

The distribution of the maximum F-to-enter under the null hypothesis for a particular case can be approximated using Monte Carlo methods. A crude way to do this would be to generate artificial values of the predictand, Y, using

$$Y = X_p\beta_p + \epsilon,$$

where X_p is the $n \times p$ matrix of actual values of the p selected predictors, β_p is an arbitrary vector of regression coefficients (zero values are the obvious choices) and the residuals, ϵ, are generated from a normal distribution with arbitrary non-zero variance. The p selected predictors are then forced into the regression and the maximum F-to-enter is calculated. The whole process is repeated say 1000 times to estimate the required distribution. The amount of computation can be greatly reduced by using a reduction to a $(k-p)$-dimensional space (or $(n-p)$ if the number of observations is less than the number of predictors) orthogonal to the space of the p selected predictors.

An alternative method involves augmenting the set of predictors with dummy variables whose values are produced from a random number generator. When the first of these artificial variables is selected, it is assumed that there is no further useful information in the remaining predictors. Suppose that we have reached that stage in a selection procedure (though in practice we would have no way of knowing this), and that there are 10 remaining real predictors plus one artificial one. The chance that the artificial predictor will be selected next is then only 1 in 11. For this method to be useful we therefore need a moderate number of artificial predictors to ensure that selection stops at about the right place, say about the same number of artificial predictors as we have real ones. This immediately makes the method less attractive as the amount of computation required increases rapidly with the number of predictors.

Table 2 shows the residual sums of squares for the five best-fitting subsets of each size for our three data sets with the following numbers of predictors: STEAM (9 + 9), DETROIT (11 + 11) and PLANES (14 + 11), where each pair contains the number of real predictors followed by the number of added artificial predictors. The asterisks in Table 2 indicate the number of artificial predictors in each subset. We note that in the case of the STEAM data, the closeness of the best-fitting subset of three predictors including an artificial predictor, to the best-fitting one, casts some doubt as to whether there is useful information after the first two predictors have been chosen. Similarly there must be considerable doubt in the case of the best-fitting subset of four predictors for the DETROIT data.

A permutation test was proposed by Forsythe *et al.* (1973) especially for the case in which there are more predictors than observations. Suppose that we have $k > p$ available predictors, and that the true model is

$$Y = X_A\beta_A + \epsilon,$$

where X_A is an $n \times (p+1)$ matrix consisting of the values of just p of the predictors and a column of 1's, β_A is a $(p+1) \times 1$ vector of regression coefficients, and the true residuals, ϵ, have variance σ^2. Let X_B be an $n \times (k-p)$ matrix containing the values of the remaining predictors. Form an orthogonal reduction of the kind:

$$(X_A, X_B) = (Q_A, Q_B)R, \tag{3.3}$$

where the columns of Q_A and Q_B are mutually orthogonal and normalized, with Q_A spanning the space of X_A, and Q_B spanning that part of the space of X_B which is orthogonal to X_A. The matrix R is usually upper triangular in regression calculations, but that is not essential for our derivation.

TABLE 2

Residual sums of squares for the five best-fitting subsets of each size for three data subsets with artificial variables added

No. of predictors	*Data set*		
	STEAM	DETROIT	PLANES
2	8.98	33.83	10.80
	9.63	44.77	11.12
	9.78	54.46	11.17
	15.39*	55.49	11.45
	15.60	62.46	11.48
3	7.34	6.77	7.56
	7.68	21.19	7.58
	7.81*	23.05	7.98
	8.29*	23.51	7.98
	8.29*	25.01*	8.00
4	6.41**	3.79	5.79
	6.72*	4.08*	6.20
	6.80	4.58	6.25
	6.93	5.24	6.49
	7.02	5.38*	6.57
5			4.97*
			5.15
			5.15
			5.27*
			5.33*

Forsythe *et al.* used Householder reductions to achieve the factorization (3.3), but several other methods are widely used. Applying the same reduction to the vector of values of Y, we obtain:

$$\begin{bmatrix} Q_A' \\ Q_B' \end{bmatrix} Y = \begin{bmatrix} I \\ 0 \end{bmatrix} R\beta + \begin{bmatrix} Q_A'\epsilon \\ Q_B'\epsilon \end{bmatrix},$$

where I is a $p \times p$ identity matrix and 0 is a $(k-p) \times p$ matrix of zeros. The last $(k-p)$ projections are then equal to $Q_B'\epsilon$, a vector whose values have zero expected value and covariance matrix:

$$E(Q_B' YY' Q_B) = \sigma^2 I.$$

That is, the projections in $Q_B'Y$ are uncorrelated and all have the same variance (unlike least-squares residuals). If the orthogonal reduction method used is that of planar rotations, then the projections have been shown by Farebrother (1978) to be identical with the "recursive" residuals of Brown *et al.* (1975), though the planar rotation method is a much more efficient way of calculating them.

It is wrongly assumed by Forsythe *et al.* that as these projections all have the same mean and variance, they also have the same distribution. This only applies if the true residuals, ϵ, have a normal distribution.

Suppose we are at the stage at which p predictors have been selected. We can find that predictor from those remaining which gives the largest F-to-enter. If the projections are exchangeable then the last $(k-p)$ projections can be permuted and the reduction in residual sum of squares for each one of the remaining $(k-p)$ variables, if it were to be selected next, can be calculated for each permutation. This can be repeated say 1000 times to find the number of times that the original maximum F-to-enter is exceeded. If there is no further useful information from the remaining predictors then the number of exceedances is equally likely to be any of the integers 0, 1, . . ., 1000.

In computation, the only difference from the maximum F-to-enter test is that the projections are permuted in this test; they are generated afresh for each case in the maximum F-to-enter test. The permutation test is usually much faster.

The main weakness of this test is that a large number of exceedances can occur when two (or more) remaining predictors can improve the fit substantially and are competing for selection next. Table 3 shows the results from applying this test to our three sets of data using 1000 permutations in each case. The selected subsets in this case are those picked using forward selection.

TABLE 3

Numbers of times out of 1000 that the maximum reduction in residual sum of squares was exceeded in a permutation test

	Number of previously selected predictors					
Data set	0	1	2	3	4	5
STEAM	17	80	711	590		
DETROIT	13	10	269	378	447	101
PLANES	27	485	311	77	739	

If we have fitted a linear model containing p out of k available predictors, then the lack-of-fit statistic is

$$LF = \frac{(\mathrm{RSS}_p - \mathrm{RSS}_k)/(k-p)}{\mathrm{RSS}_k/(n-k-1)} \quad ,$$

provided that $n > k + 1$. If the subset of p variables had been chosen *a priori* then this statistic has an F-distribution if the true regression coefficients of the remaining $(k-p)$ predictors are all zero, and of course subject to the usual conditions of independence, normality and homogeneity of variance of the true residuals. If $(k-p)$ is large and there is only one further useful predictor to be found, then its contribution to RSS_p tends to be "swamped" by the rest of the $(k-p)$ so that the test lacks power compared with the maximum F-to-enter. However, if there are two or more remaining variables then, as RSS_k rather than RSS_{p+1} is used in the denominator, the lack-of-fit statistic is more powerful than the maximum F-to-enter. If the value of LF is appreciably less than 1.0 it can indicate that "over-fitting" has occurred in the selection of the first p predictors.

Spjøtvoll's (1972a) test is of whether one subset fits better than another. The measure of goodness-of-fit which he uses is the sum of squares of differences between the expected values of the predictand and the predictions. It will be assumed that the elements of Y are independently and normally distributed about an expected value vector $\eta(X)$. No assumptions will be made at this stage about the functional form of the relationship between η and X. Consider a subset A of the predictors and let X_A be a matrix containing the columns of X for the predictors in A. Taking

$$\beta_A = (X_A' X_A)^{-1} X_A' \eta,$$

Spjøtvoll's measure of goodness-of-fit is

$$(\eta - X_A \beta_A)' (\eta - X_A \beta_A) = \eta' \eta - \eta' X_A (X_A' X_A)^{-1} X_A' \eta. \tag{3.4}$$

An alternative argument is to use the predicted values $X_A \hat{\beta}_A$ instead of their expected values $X_A \beta_A$. This adds an extra term

$$\sigma^2 \cdot \mathrm{trace}\ [X_A (X_A' X_A)^{-1} X_A'] \tag{3.5}$$

to expression (3.4) where σ^2 is the variance of the predictand. Provided that there are no linear dependencies amongst the predictors in A, this is equal to $(p_A + 1)\sigma^2$ where p_A is the number of predictors in A.

As the first term on the right-hand side of (3.4) does not vary when we change the subset A, Spjϕtvoll chose to use as his measure of goodness-of-fit

$$\eta' X_A (X_A' X_A)^{-1} X_A' \eta \tag{3.6a}$$

which is similar to a regression sum of squares except that it uses η instead of Y. If prediction is our objective, then it would seem to be more appropriate to add the extra term given in (3.5) to give the quantity:

$$\eta' X_A (X_A' X_A)^{-1} X_A' \eta - (p_A + 1)\sigma^2, \tag{3.6b}$$

A larger value of the measure for one subset than the other implies a better fit. For two subsets A and B we want then to make inferences about

$$\eta' \, [X_A (X_A' X_A)^{-1} X_A' - X_B (X_B' X_B)^{-1} X_B'] \, \eta - (p_A - p_B)\sigma^2, \tag{3.7}$$

where the last term was not used by Spjϕtvoll.

Spjϕtvoll's test is a Scheffé-type test. Suppose that $Y = X\beta + \epsilon$, where the true residuals, ϵ, are independently and normally distributed with variance σ^2, so that $\eta = X\beta$. That is that we have the true model if we use all of the available predictors, though an unknown number of the β's may be zero. Using $\hat{\beta}$ to denote the least-squares estimate of β, then

$$\Pr \, [(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \leqslant (k+1) \, s^2 F_{\alpha, k+1, n-k-1}] = 1 - \alpha \tag{3.8}$$

where $F_{\alpha, k+1, n-k-1}$ is the upper 100α per cent point of the F-distribution with $(k+1)$ and $(n-k-1)$ degrees of freedom, and s^2 is the usual estimate of the residual variance, i.e. $\text{RSS}_k/(n-k-1)$.

If the linear model in all the predictors does not include the true model, i.e. $\eta \neq X\beta$, then $(n-k-1)s^2/\sigma^2$ has a non-central χ^2 distribution and the probability in (3.8) is greater than $(1-\alpha)$, as shown by Scheffé (1959, p. 136-137).

Substituting $\eta = X\beta$ in (3.7) and neglecting its last term for the moment, we want confidence limits for

$$\beta' X' \, [X_A (X_A' X_A)^{-1} X_A' - X_B (X_B' X_B)^{-1} X_B'] \, X\beta = \beta' C \beta \tag{3.9}$$

when β satisfies the condition on the left-hand side of (3.8). If $\eta = X\beta$ is not the true relationship, that part of the space of η which is orthogonal to X is also orthogonal to any subset, so that the projections in (3.7) of η onto the space of the subsets are identical with the projections of $X\beta$ onto this space. Thus we want to find the largest and smallest values of the quadratic form $\beta' C \beta$ given that the closeness of the unknown β to the known $\hat{\beta}$ is described by the ellipsoid defined in (3.8). An equivalent problem posed by C. R. Rao was solved by Forsythe and Golub (1965). Spjϕtvoll (1972b) gives a more detailed derivation for this particular problem. The formula for the $100(1-\alpha)$ per cent confidence limits are contained in Theorem 1 on p. 1079 of Spjϕtvoll (1972a). The present author has derived an efficient computational method and FORTRAN code for finding these limits can be provided on request.

An important feature of Spjϕtvoll's confidence limits is that they are simultaneous limits for comparisons between all possible pairs of different subsets for the same data set. The confidence limits tend to be conservative, as is usual with Scheffé-type limits. The Spjϕtvoll procedure requires very little computation compared with several of the other tests which have been described. Spjϕtvoll's method can be used, as one referee has suggested, to find all subsets which do not differ significantly from one another, though the number of such subsets will often be prohibitively large.

Table 4 shows the upper and lower 90 per cent confidence limits, A_1 and A_2, for Spjϕtvoll's measure of goodness-of-fit for a few selected pairs of subsets of interest for the STEAM data.

TABLE 4

Upper and lower 90 per cent confidence limits, A_1 and A_2, for Spjøtvoll's measure for comparing the goodness-of-fit of pairs of subsets applied to the STEAM data

Subset A	*Subset B*	$\text{RSS}_B - \text{RSS}_A$	A_1	A_2
7	6	19.4	3.7	39.6
7	5	27.2	−7.1	66.9
7	3	31.2	8.1	62.0
7	8	35.7	11.6	68.7
7	1, 7	−9.3	−30.4	−0.34
7	4, 5, 7	−10.9	−33.3	−0.69
1, 7	4, 5, 7	−1.6	−14.3	10.1
1, 7	1-9 incl.	−4.1	−20.1	0.0

The differences in the residual sums of squares for subsets A and B are given for comparison with the confidence limits. This difference must lie between t'ıe limits A_1 and A_2. We note that the best-fitting subset of two predictors (numbers 1 and 7) fits significantly better at the 10 per cent level than the best-fitting single predictor (number 7), though the reductions from adding further predictors are not significant. The single predictor comparisons are interesting. Predictor number 7 gives a significantly better fit than the second, fourth and fifth-best predictors, but not significantly better than the third-best. The product-moment correlations between predictor number 7 and these other four predictors and the range of the confidence limits are:

Predictor	6	5	3	8
Correlation	−0.86	−0.21	−0.62	−0.54
Range	35.9	74.0	53.8	57.1

Where there are high correlations, positive or negative, between the predictors in two subsets, they span almost the same space and so narrow confidence limits can be expected. In this example, we see a clear relationship between the correlations and the range of the confidence limits.

If certain predictors are to be forced into all subsets then slightly narrower confidence limits can be obtained, as explained by Spjøtvoll on p. 1085 of his paper. If r predictors are to be forced in, then (3.8) can be replaced with an equivalent statement in terms of the components of the remaining $(k - r)$ predictors which are orthogonal to the r predictors forced in. The "$k + 1$" which multiplies s^2 is then replaced with $(k + 1 - r)$, as also is the $(k + 1)$ for the numerator degrees of freedom for F. In our example we have treated the constant as one such predictor.

Some special cases of testing between subsets have also been considered by Aitkin (1974) and Tarone (1976), while Borowiak (1981) appears to have derived a similar result to Spjøtvoll's but for the case in which σ^2 is assumed known.

The most-widely used statistic for testing whether there is any predictive value in a selected subset is the "coefficient of determination" or "multiple R^2" defined as $(\text{RSS}_0 - \text{RSS}_p)/\text{RSS}_0$. An interesting example of the use of R^2 is contained in a paper by McQuilkin (1976). Subset regression methods are used to predict tree heights as a function of soil and site conditions. Data were obtained on 50 predictors, many of which were polynomial or interaction terms, for trees in 81 plots along a ridge top. The 81 plots were divided into two groups with every third plot along the ridge, from a random start amongst the first three, going into the second group. Using an exhaustive search procedure on the data from the first group of 54 plots, a subset of 8 predictors was selected for predicting the average height of trees in a plot. The value of R^2 was 0.66. Using least-squares estimates of the regression coefficients, the average heights were predicted for the other 27 plots. The value of R^2 for the regression of the actual and predicted heights was found to be only 0.01. I am grateful to Ken Berk for referring me to this example.

There has been a number of empirical tabulations of the distribution of R^2 when Y is normally distributed and independent of the predictors. Diehr and Hoflin (1974), and Lawrence *et al.* (1975) have generated the distribution using respectively exhaustive research and forward selection for uncorrelated normally distributed predictors. Zurndorfer and Glahn (1977) and Rencher and Pun (1980) have looked at the case of correlated predictors using forward selection and the Efroymson algorithm respectively. It appears from both of these studies that higher values for R^2 result when the predictors are uncorrelated. Wilkinson and Dallal (1981) considered the same case as Lawrence *et al.*, but gave their tables in terms of the number of remaining predictors, $(k-p)$, and the F-to-enter value used as the stopping rule.

Zirphile (1975) attempted to use extreme-value theory to derive the distribution of R^2. Rencher and Pun developed this idea further and obtained a formula for the upper percentiles of R^2 with two parameters which they adjusted to fit their tables.

When the number of observations exceeds the number of available predictors, so that a valid unbiassed estimate of the residual variance is available, the use of Spjøtvoll's test is adequate for most practical purposes and is easy to apply. The coefficient of determination is a popular measure and can be used as an alternative to Spjøtvoll's test for testing whether there is any predictive value in a selected subset, using empirical tables or the formula given by Rencher and Pun. When there are more predictors than observations, the only valid test known to the author is that of adding artificial variables to the set of available predictors; this does not though provide any way of testing whether one subset fits better than another.

4. ESTIMATION

If least squares is used to estimate regression coefficients for a subset of predictors when that subset has been selected using the same data, then the regression coefficients will be biassed. This has been known for a long while, e.g. Miller (1962), yet it is still almost the only method of estimation used. Biassed estimators are widely used in many areas of statistics, and provided that the properties of the estimators are known, this usually causes little concern. For instance, few statisticians use unbiassed estimates of standard deviations. In this section, this bias will be examined in a simple case, and several methods of estimation will be considered.

Let $X=(X_A, X_B)$ be a sub-division of the set of available predictors into two subsets A and B, and let $\beta=(\beta_A, \beta_B)$ be the corresponding sub-division of the regression coefficients. If the sub-division has been carried out independently of the observed values of Y, then the expected values of the least-squares coefficients, b_A, for subset A are

$$E(b_A)=\beta_A+(X_A'X_A)^{-1}X_A'X_B\beta_B. \tag{4.1}$$

This is the vector of unconditional expected values of the regression coefficients for subset A. The second term on the right-hand side of (4.1) could be considered the bias in estimating β_A arising from the omission of the predictors in subset B. In many cases in which subset regression is used, all or most of the predictors are random variables, and it is then convenient to think of the model in subset A predictors as

$$Y=X_A\beta_A^*+e(X)+[\langle I-X_A(X_A'X_A)^{-1}X_A'\rangle X_B\beta_B+\epsilon], \tag{4.2}$$

where β_A^* is given by (4.1), and that part of $X_B\beta_B$ which is orthogonal to X_A is treated as additional random variation augmenting ϵ. This is the model which would be used if the values of the variables in subset B were not available.

If subset A is chosen conditional upon the subset fitting better than certain others, then $E(b_A)$ will in general be different from (4.1) if selection and least-squares estimation are from the same data. This difference will be called the selection bias. It may be convenient to think of this selection bias as being due to two causes, the, first is competition for selection amongst several subsets containing the same numbers of predictors, and the second arises from the stopping rule which is applied to decide the number of predictors to use. Thus we now have the three kinds of bias:

(i) omission bias, equal to the second term on the right-hand side of (4.1),

(ii) competition bias, in choosing between subsets of the same size, and

(iii) stopping-rule bias, in choosing the number of predictors to use.

Stopping-rule bias has been considered in detail by Kennedy and Bancroft (1971) for the case in which the subsets of each size are pre-specified, as for instance is usually the case in fitting polynomial regressions of progressively higher degree or in adding the next higher lag in fitting auto-regressive models. Copas (1983, Section 6) considered the bias in the case of orthogonal predictors; the bias in that case was a combination of competition and stopping-rule bias. These are special cases of an area of statistics known as pre-test estimation, see, for example, Bancroft and Han (1977) or Judge and Bock (1978).

Let us look at the simple case of two competing predictors, and suppose that it has been decided to use only one of them. Suppose that the true model is

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{4.3}$$

where the residuals, ϵ, have zero mean and $E(\epsilon^2) = \sigma^2$. Later we will also need to make assumptions about the distribution of the residuals. The least-squares estimate, b_1, of the regression coefficient for the simple regression of Y upon X_1 is then

$$b_1 = X_1' Y / X_1' X_1$$

and hence

$$E(b_1) = \beta_1 + \beta_2 X_1' X_2 / X_1' X_1$$

$$= \beta_1^* \text{ say,}$$

with a similar definition for β_2^*. Note that these are the expected values over all samples, no selection has been considered so far. The difference between β_1^* and β_1, and similarly that between β_2^* and β_2, is what we called earlier the *omission* bias.

Variable X_1 is selected when it gives the smaller RSS, or equivalently, when it gives the larger regression sum of squares. That is, when

$$X_1' X_1 b_1^2 > X_2' X_2 b_2^2. \tag{4.4}$$

If we let $f(b_1, b_2)$ denote the joint probability density of b_1 and b_2, then the expected value of b_1 when variable X_1 is selected is

$$E(b_1 \mid X_1 \text{ selected}) = \frac{\int_R \int b_1 f(b_1, b_2)\, db_1\, db_2}{\int_R \int f(b_1, b_2)\, db_1\, db_2}, \tag{4.5}$$

where the region R in the (b_1, b_2)-space is that in which condition (4.4) is satisfied. The denominator of the right-hand side of (4.5) is the probability that variable X_1 is selected. The region R can be re-expressed as that in which $|b_1| > C|b_2|$ where $C = (X_2' X_2 / X_1' X_1)^{\frac{1}{2}}$. As the boundaries of this region are straight lines, it is relatively straightforward to evaluate (4.5) numerically for any assumed distribution of b_1 and b_2, given the sample values of $X_1' X_1$ and $X_2' X_2$. Similarly, by replacing the b_1 in the numerator of the right-hand side of (4.5) with b_1^r, we can calculate the rth moment of b_1 when X_1 is selected.

As b_1 and b_2 are both linear, in the residuals, ϵ, it is feasible to calculate $f(b_1, b_2)$ for any distribution of the residuals. If the residuals have a distribution which is close to normal then, by the Central Limit Theorem, we can expect the distribution of b_1 and b_2 to be closer to normal, particularly if the sample size is large. The results which follow are for the normal distribution.

Given the values of X_1 and X_2, the covariance matrix of b_1, b_2 is

$$V = \sigma^2 \begin{bmatrix} (X_1'X_1)^{-1} & X_1'X_2(X_1'X_1)^{-1}(X_2'X_2)^{-1} \\ X_1'X_2(X_1'X_1)^{-1}(X_2'X_2)^{-1} & (X_2'X_2)^{-1} \end{bmatrix}.$$

Without loss of generality, it will be assumed that X_1 and X_2 have been scaled so that

$$V = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{where } \rho = X_1'X_2.$$

The joint probability density of b_1 and b_2 is then

$$f(b_1, b_2) = \frac{\exp\{-\tfrac{1}{2}(b-\beta^*)' V^{-1}(b-\beta^*)\}}{2\pi\sigma^2(1-\rho^2)^{\frac{1}{2}}}.$$

Let $\mu(b_2) = \beta_1^* + \rho(b_2 - \beta_2^*)$, then we need to evaluate integrals of the form

$$I_r = \int_{-\infty}^{\infty} \frac{\exp[-(b_2-\beta_2^*)^2/2\sigma^2]}{(2\pi\sigma^2)^{\frac{1}{2}}} \cdot \int_{R(b_2)} \frac{b_1^r \exp\{-[b_1-\mu(b_2)]^2/[2\sigma^2(1-\rho^2)]\}}{\{2\pi\sigma^2(1-\rho^2)\}^{\frac{1}{2}}}\, db_1\, db_2, \quad (4.6)$$

where the regions of integration for the inner integral are $b_1 > |b_2|$ and $b_1 < -|b_2|$. The inner integral can be evaluated easily for low moments. Numerical integration can then be used to determine I_r. Unfortunately, none of the derivatives of the kernel of (4.6) is continuous at $b_2 = 0$, so that normal Hermite integration cannot be used. However, the kernel is well behaved on each side of $b_2 = 0$ so that integration in two parts can easily be carried out using half-Hermite integration. Tables for half-Hermite integration have been given by Steen *et al.* (1969), and by Kahaner *et al.* (1982).

Table 5 contains some values of the mean and standard deviation of b_1 when variable X_1 is selected. In this table the unconditional expected value of b_1 is held at 1.0.

TABLE 5

Values of the expected value, $E(b_1 \mid sel.)$, and standard deviation, st. dev. $(b_1 \mid sel.)$, of b_1 when variable X_1 is selected, with $\beta_1^* = 1.0$

		$\sigma = 0.3$		$\sigma = 0.5$	
	β_2^*	$E(b_1 \mid sel.)$	*St. dev.* $(b_1 \mid sel.)$	$E(b_1 \mid sel.)$	*St. dev.* $(b_1 \mid sel.)$
$\rho = -0.6$	0.0	1.02	0.28	1.11	0.43
	0.5	1.08	0.25	1.21	0.39
	1.0	1.21	0.21	1.36	0.35
	1.5	1.39	0.18	1.53	0.32
	2.0	1.60	0 16	1.72	0.30
$\rho = 0.0$	0.0	1.01	0.29	1.10	0.45
	0.5	1.05	0.28	1.15	0.44
	1.0	1.17	0.25	1.28	0.42
	1.5	1.35	0.23	1.46	0.40
	2.0	1.57	0.22	1.66	0.38
$\rho = 0.6$	0.0	1.02	0.28	1.11	0.43
	0.5	1.01	0.29	1.09	0.46
	1.0	1.11	0.28	1.17	0.48
	1.5	1.30	0.27	1.34	0.51
	2.0	1.53	0.27	1.52	0.58

Fig. 1 is intended to give a geometric interpretation of selection bias. The ellipses are for two different cases, and are ellipses of constant probability density in (b_1, b_2) such that most pairs of values of (b_1, b_2) are contained within them. For this figure, it is assumed that X_1 and X_2

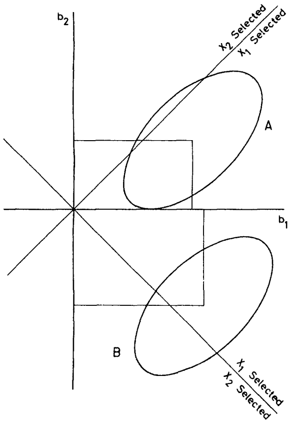

Fig. 1. Illustrating regions for selection in the space of the regression coefficients when one variable out of two is to be selected.

have both been scaled to unit length so that the regions in which X_1 and X_2 are selected are bounded by lines at 45 degrees to the axes. Thus X_1 is selected if (b_1, b_2) is in regions to the left and right of the origin, and X_2 is selected in the top or bottom regions.

Ellipse A represents a case in which b_1 is positive and b_2 is usually positive. The thin horizontal and vertical lines running from the centroid of the ellipse are at the unconditional expected values of b_1 and b_2. When X_2 is selected, (b_1, b_2) is in the small sliver to the top left of the ellipse or just above it. Most of the sliver is above the expected value of b_2, so that b_2 is biassed substantially in those rare cases in which it is selected. As the few cases in which X_1 is not selected give values of b_1 less than its expected value, b_1 is biassed slightly on the high side when X_1 is selected.

Ellipse B represents a case in which the principal axis of the ellipse is perpendicular to the nearest selection boundary. In this case, far more of the ellipse is on the "wrong" side of the boundary and the biasses in both b_1 and b_2, when their corresponding variables are selected, are relatively large.

In both case A and case B, the standard deviations of b_1 and b_2 when their variables are selected, are less than the unconditional standard deviations. This applies until the ellipses containing most of the joint distribution include the origin. It can be seen that when β_1^* and β_2^* have the same signs and are well away from the origin, the biasses are smallest when $\rho \gg 0$ and largest when $\rho \ll 0$. The case $\rho = 0$, that is when the predictor variables are orthogonal, gives an intermediate amount of bias. The popular belief that orthogonality gives protection against selection bias is fallacious; highly correlated variables can give more protection.

The above derivations have been of the properties of the regression coefficients. A similar exercise can be carried out using the joint distribution of the RSS's for the two variables, to find the distribution of the minimum RSS. This is somewhat simpler as both RSS's must be positive or zero, and the boundary is simply the straight line at which the two RSS's are equal.

The case of competition between two variables has been studied in greater detail by Christenson (1982) who used a cost function decision rule to pick 0, 1 or 2 variables.

To extend the above theory to the case of k competing predictors when only one is to be selected is not difficult; the condition (4.4) which defines the region of selection is replaced with $(k-1)$ such conditions, and the bivariate integration in (4.5) becomes a k-dimensional integration. However, once we try to extend it to the case of the best-fitting two or more predictors out of k, the conditions in (4.4) involve general quadratic forms so that the limits of the inner integrals are no longer linear but are in terms of the solutions of quadratic equations.

If the predictors are orthogonal, as for instance when the data are from a designed experiment or when the user has constructed orthogonal variables from which to select, then we can easily derive upper limits for the competition bias. If we scale all the predictors to have unit length, then the worst case is when the expected values of the regression coefficients are all the same in absolute value. If all of the regression coefficients have expected value equal to $\pm\beta$ with sample standard deviation equal to $\sigma\,(\sigma \ll \beta)$, then if we pick just one predictor, that with the largest regression coefficient in absolute value, the expected value of the absolute value of its regression coefficient is $\beta + \xi_1 \sigma$, where ξ_1 is the first-order statistic for a random sample of k values from the standard normal distribution, where k is the number of predictors available for selection. If we pick the three predictors which give the best fit to a set of data, then the bias in the absolute values of the regression coefficients will have expected value equal to $\sigma(\xi_1 + \xi_2 + \xi_3)/3$, where ξ_i is the ith normal order statistic. Thus, if we have say 25 available predictors, the bias in the regression coefficient of a single selected predictor will be about 1.97 standard deviations.

The order-statistic argument gives only a rough guide to the likely size of the competition bias in general, though it does give an upper limit when the predictors are orthogonal. The competition bias can be higher than the order-statistic limit for correlated predictors. In the author's experience, competition biasses of over two standard deviations are fairly common in real life problems, particularly when an exhaustive search has been used to select the chosen subset of predictors. If forward selection is used, the competition bias is usually smaller, but then the selected subsets may be much inferior to those from an exhaustive search.

Simulation could yield a useful empirical formula for the bias in the regression coefficients. As most of the bias results from the competition for selection, some index of the extent of competition would have to be the main ingredient in such a formula.

What can be done to overcome this bias? Amongst possible methods are:

(i) Use only half of the available data to select the predictors, and the other half for estimation. If the halves are chosen randomly, the regression coefficients will be unbiassed. By only using part of the data, the chances of finding the best-fitting subset are reduced. In many situations, this is unimportant; we merely require a subset which gives good predictions, though the smaller the set of data used, the more likely we are to pick a poor subset. In many fields, sample sizes are not large enough to make this method practical.

(ii) A jack-knife method. The results for competition between two variables, and the order-statistic argument for orthogonal predictors both suggest that the competition bias is of order $n^{-\frac{1}{2}}$, where n is the sample size. If $b(n)$ is the vector of sample regression coefficients for the selected variables using n observations, the estimate

$$\frac{n^{\frac{1}{2}} b(n) - (n-r)^{\frac{1}{2}} b(n-r)}{n^{\frac{1}{2}} - (n-r)^{\frac{1}{2}}},$$

where r is the number of omitted observations, removes a bias of order $n^{-\frac{1}{2}}$. In using this method, r could be taken as say 10 per cent of n. The estimate $b(n-r)$ is used only if the same selection and stopping rule as were used for the full data set selects the same subset of variables. Thus, in say 100 sets of $(n-r)$ out of n observations, it may be found that only 70 of them result in the selection of the original subset. No investigation has been made of properties of this method. The idea depends very heavily upon the assumption that the bias is of $O(n^{-\frac{1}{2}})$. If this is not so, the use of this version of the jack-knife could make the bias problem worse. Several readers of drafts of this paper have also suggested the use of bootstrap methods.

(iii) Shrunken estimators. Either ridge regression or a Stein-type estimator will often reduce the bias of most of the regression coefficients, for suitable values of the parameter controlling shrinkage. For two hybrid estimators which combine some of the properties of a ridge estimator with subset selection, see Hemmerle and Carey (1983).

(iv) Use Monte Carlo methods to estimate the bias, using estimates of the regression coefficients as if they were population values. The bias can then be subtracted from the estimates. This is a straightforward method which may be adequate in many cases. The first set of corrected estimates can be used as new population values, and the process repeated iteratively.

(v) Maximum likelihood. Given n observations of Y and of all k available predictors, and assuming normality, the unconditional likelihood is

$$\prod_{i=1}^{n} \phi[(y_i - \Sigma\, \beta_j x_{ij})/\sigma],$$

where ϕ is the standard normal probability density. We want to estimate regression coefficients when a particular subset has been selected by some procedure (stepwise regression, exhaustive search, etc.). Many vectors of values of Y are then impossible as they lead to the selection of other subsets. For the population of all vectors of n values of the Y-variable such that our subset is selected, the likelihood for the actual sample is

$$\frac{\prod_{i=1}^{n} \phi[(y_i - \Sigma\beta_j x_{ij})/\sigma]}{\int_R \ldots \int (\text{above density})\, dy_1 \ldots dy_n},$$

where the region R is that part of the space of Y in which our subset is selected; the likelihood is zero outside of this region. The denominator is the probability of selection of our subset, which is a function of the selection procedure and stopping rule used. This likelihood involves all k predictors, not just those in the selected subset. Note that the boundaries of R are not functions of the regression coefficients, so that small-sample likelihood theory applies.

Substituting for ϕ, the logarithm of the conditional likelihood is then

$$-(n/2)\log_e(2\pi\sigma^2) - (2\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \Sigma\beta_j x_{ij})^2 - \log_e(P), \tag{4.7}$$

where P is the probability of selection of our subset. Maximizing (4.7) yields estimates of the regression coefficients for all k available predictors, and by projection onto the space of our subset of predictors, regression coefficients for those variables can be obtained.

Clearly, the estimation of the probability of selection is the main obstacle to the use of this method. A number of methods have been tried, and the author believes that he now has a feasible method. This will be reported elsewhere when its development is complete. Preliminary results indicate a substantial reduction in competition bias. No attempt has been made to allow for stopping-rule bias at this stage.

5. MEAN SQUARED ERRORS OF PREDICTION AND STOPPING RULES

Let us write

$$Y = X_A\beta_A + e(X_A) + \epsilon, \tag{5.1}$$

where the subscript A denotes a particular subset of variables, X_A and β_A are the design matrix and linear regression coefficients for this subset, $e(X_A)$ is the deterministic error in this model (e.g. from assuming a linear instead of a non-linear model, or from neglecting interactions) and ϵ is the residual variation, some of which is from omitted variables. The deterministic error will be defined to be orthogonal to X_A. Further assume that the residuals have zero mean, constant variance σ_A^2, and are independent of each of the predictors in subset A.

Let x_A be a vector of values of the selected variables for which we want to obtain a prediction. If b_A is our sample estimate (not necessarily using least squares) of β_A, then suppose we predict Y as

$$\hat{Y}(x_A) = x_A' b_A.$$

The expected value of the squared prediction error is then

$$E[Y - \hat{Y}(x_A)]^2 = \underset{(A)}{x_A' [(\text{bias})(\text{bias})'} + \underset{(B)}{V(b_A)] x_A} + \underset{(C)}{2x_A'(\text{bias}) e(x_A)} + \underset{(D)}{e^2(x_A)} + \underset{(E)}{\sigma_A^2}, \tag{5.2}$$

where "bias" = $E(b_A - \beta_A)$, $V(b_A)$ is the covariance matrix for the sample regression coefficients and $e(x_A)$ is the deterministic error at the point x_A.

In most derivations of mean squared errors of prediction (MSEP), only the terms (B) and (E) are considered. The values of these terms are usually averaged over either the set of values of x_A in the design matrix, or over a multivariate normal distribution with the same covariance matrix as for the design matrix. Either gives a simple result for the expected value of (B). By assuming that the sample residual variance estimate, say s_A^2, provides an unbiassed estimate of σ_A^2, a stopping rule is obtained by finding the size of subset which minimizes an estimate of (5.2). This is the basic method which yields Mallows' C_p and several similar quantities (see, for example, Thompson, 1978). Alternatively, likelihood arguments can be used leading to Akaike's Information Criterion (AIC) and variants upon it.

Mallows' C_p is

$$C_p = \frac{\text{RSS}_p}{s^2} + 2p - n,$$

where $s^2 = \text{RSS}_k/(n-k-1)$. At its minimum we have that $C_p < C_{p+1}$, and hence

$$\frac{\text{RSS}_p - \text{RSS}_{p+1}}{s^2} < 2.$$

This quantity will often be nearly equal to the F-to-enter for the $(p+1)$st variable, so that minimizing Mallows' C_p is roughly equivalent to using an F-to-enter of 2.0, or equivalently a t-value of about 1.4, as the stopping rule. Similar calculations for the AIC show that at its maximum,

$$\frac{\text{RSS}_p - \text{RSS}_{p+1}}{\text{RSS}_{p+1}/(n-p-2)} \leqslant (n-p-2)(e^{2/n} - 1).$$

Providing that $n \gg 2$, this means that maximizing the AIC is roughly equivalent to using an F-to-enter of $[2 - 2(p+2)/n]$. When Mallows' C_p or the AIC is plotted against p, there are often several local minima so that minimizing or maximizing such quantities cannot be exactly equated to the use of an F-to-enter test. Bendel and Afifi (1977) show that minimizing the AIC is close to the optimal stopping rule for prediction *when unbiassed estimates of regression coefficients and of the residual variance are available*.

As we have seen, the competition bias can easily be of the order of two standard errors when selection and estimation are from the same data. This means that term (A) in (5.2) can be of the order of four times the size of term (B). The usual residual variance estimate is an estimate of (D) + (E) and will usually be biassed on the low side. As the deterministic error has been defined to be orthogonal to X_A, the term (C) will usually be small except for extrapolation. Hence the vast literature on stopping rules (see Breiman and Freedman, 1983 or Kohn, 1983 for recent references) is an irrelevant academic exercise until the problems of estimation have been overcome.

To illustrate the effect of neglecting the estimation bias, the PLANES data have been used.

The sample estimates of the regression coefficients and the residual variance were taken as population values for the generation of 250 artificial data sets, each consisting of the same 31 sets of predictors with random normal noise added. For each set of artificial data, the best-fitting subsets of all sizes were found. For each best-fitting subset, the residual variance, σ_A^2 was estimated using $s_A^2 = \text{RSS}_p/(n-p-1)$ where p = the number of predictors in the subset excluding the constant. The MSEP for that subset was estimated for the values of x_A in the original design matrix, ignoring estimation bias, using

$$\text{MSEP (false)} = [1 + (p+1)/n]\, s_A^2,$$

and compared with an estimate of (5.2) obtained by using the known true value of σ_A^2, and estimating (A) + (B) using $(b_A - \beta_A)'(b_A - \beta_A)$. Fig. 2 shows the outcome. In this case, the true MSEP is a minimum when all 14 predictors are included. The horizontal line in Fig. 2 is at the level of the residual variance for the full model (0.221). As the false MSEP almost reaches this line, it means that the estimates s_A^2 are less than this for some subsets. The numbers of different best-fitting subsets of each size for the 250 data sets were:

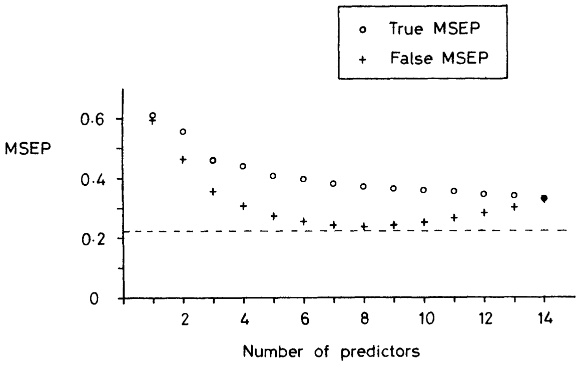

Fig. 2. True and false MSEP against the number of predictors, for the PLANES data set.

Size of subset	1	2	3	4	5	6	7	8	9	10	11	12	13
No. of different "best" subsets	6	16	32	83	97	133	148	169	169	168	136	67	14

Notice that the true and false MSEP are close together when there is little competition for selection. Even though 6 different single variables were picked, variable number 2 was selected in 233 out of the 250 data sets. The true MSEP decreases monotonically with increasing subset size in this case. It is probable that it can have a local maximum when there are a few dominant predictors which are always picked first and do not compete amongst each other, and many other less useful predictors. It can be anticipated that the true and false MSEP's will come closer relatively as sample sizes increase and the competition bias decreases.

Similar results to the above have previously been reported by Berk (1978b) and Copas (1983). Berk used two different estimates of the MSEP, both of which under-estimated the true MSEP by a factor greater than 3 in the centre of the MSEP versus p curve, for a data set with 14 predictors. Hjorth (1982) has proposed a method of estimating the true MSEP by sequentially using a subset procedure to fit a model to part of the data, estimating the next observation, then using the subset procedure again with the new observation included to predict the next point. This does not help in finding a good prediction equation, but it does give a much better idea of how bad the subset regression predictor really is.

The most important problem in subset selection is that of handling competition bias. As this bias is squared in (5.2), if the bias can only be halved, the importance of term (A) is substantially reduced, and subset selection may become competitive for prediction with either shrunken estimators or using the full model.

6. CONCLUSIONS

1. In finding best-fitting subsets, the use of replacement algorithms, particularly two-at-a-time replacement, will sometimes find much better-fitting subsets than forward selection or the Efroymson algorithm, though an exhaustive search with branch-and-bound is recommended when feasible. It is recommended that the best 10 or 20 subsets of each size, not just the best one, should be saved. The closeness of fit of these competitors gives an indication of the likely bias in least-squares regression coefficients.
2. Most classical hypothesis tests are only valid when the hypothesis has been determined *a priori*. A test due to Spjøtvoll has been described which is valid provided that there are more observations than available predictors.
3. Three sources of bias are identified, namely those due to omission, competition and the application of a stopping rule. Biases of the order of 1-2 standard errors are common in regression coefficients when the same data have been used for both model selection and estimation. This is the area of this subject which most needs further research; five possible ways of reducing the bias are suggested in the paper.
4. It is shown that the theory behind most derivations of the mean squared error of prediction, Mallows' C_p and Akaike's Information Criterion are not valid when model selection and estimation are from the same data.

REFERENCES

Aitkin, M. A. (1974) Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, **16**, 221–227.

Bancroft, T. A. and Han, C-P. (1977) Inference based on conditional specification: a note and a bibliography. *Int. Statist. Rev.*, **45**, 117–127.

Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967) The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Bendel, R. B. and Afifi, A. A. (1977) Comparison of stopping rules in forward "stepwise" regression. *J. Amer. Statist. Ass.*, **72**, 46–53.

Benedetti, J. K. and Brown, M. B. (1978) Strategies for the selection of log-linear models. *Biometrics*, **34**, 680–686.

Berk, K. N. (1978a) Comparing subset regression procedures. *Technometrics*, **20**, 1–6.

——— (1978b) Sequential PRESS, forward selection, and the full regression model. *Proc. Statist. Comput. Section, Amer. Statist. Assoc.*, 309–313.

Borowiak, D. (1981) A procedure for selecting between two regression models. *Commun. in Statist.*, **A10**, 1197–1203.

Breiman, L. and Freedman, D. (1983) How many variables should be entered in a regression equation? *J. Amer. Statist. Ass.*, **78**, 131–136.

Brown, M. B. (1976) Screening effects in multidimensional contingency tables. *Appl. Statist.*, **25**, 37–46.

Brown, R. L., Durbin, J. and Evans, J. M. (1975) Techniques for testing the constancy of regression relationships over time. *J. R. Statist. Soc.* B, **37**, 149–163.

Christenson, P. D. (1982) *Variable Selection in Multiple Regression*. Ph.D. dissertation, Iowa State University. Available from University Microfilms: Ann Arbor and London, Thesis no. 8307741.

Clarke, M. R. B. (1981) Algorithm AS 163: A Givens algorithm for moving from one linear model to another without going back to the data. *Appl. Statist.*, **30**, 198–203.

Copas, J. B. (1983) Regression, prediction and shrinkage (with Discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

Cox, D. R. and Snell, E. J. (1974) The choice of variables in observational studies. *Appl. Statist.*, **23**, 51–59.

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977) A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Ass.*, **72**, 77–106 (including discussion).

Diehr, G. and Hoflin, D. R. (1974) Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, **16**, 317–320.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979) *LINPACK Users Guide*. Soc. for Industrial and Appl. Math.: Philadelphia.

Draper, N. R., Guttman, I. and Kanemasu, H. (1971) The distribution of certain regression statistics. *Biometrika*, **58**, 295–298.

Draper, N. R., Guttman, I. and Lapczak, L. (1979) Actual rejection levels in a certain stepwise test. *Commun. in Statist.*, **A8**, 99–105.

Draper, N. R. and Smith,·H. (1981) *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Efroymson, M. A. (1960) Multiple regression analysis. In *Mathematical Methods for Digital Computers*, Vol. 1 (A. Ralston and H. S.Wilf, eds), pp. 191–203. New York: Wiley.

Elden, L. (1972) *Stepwise Regression Analysis with Orthogonal Transformations*. Unpubl. report, Mathematics Dept., Linkoping Univ., Sweden.

Farebrother, R. W. (1978) An historical note on recursive residuals. *J. R. Statist. Soc.* B, **40**, 373–375.

Fisher, J. C. (1976) Homicide in Detroit: the role of firearms. *Criminology*, **14**, 387–400.

Forsythe, A. B., Engelman, L., Jennrich, R. and May, P. R. A. (1973) A stopping rule for variable selection in multiple regression. *J. Amer. Statist. Ass.*, **68**, 75–77.

Forsythe, G. E. and Golub, G. H. (1965) On the stationary values of a second-degree polynomial on the unit sphere. *SIAM J.*, **13**, 1050–1068.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Statist. Ass.*, **76**, 817–823.

Furnival, G. M. and Wilson, R. W. (1974) Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Gabriel, K. R. and Pun, F. C. (1979) Binary prediction of weather events with several predictors. *6th Conference on Prob. & Statist. in Atmos. Sci., Amer. Meteor. Soc.*, pp. 248–253.

Garside, M. J. (1971a) Some computational procedures for the best subset problem. *Appl. Statist.*, **20**, 8–15.

——— (1971b) Algorithm AS 38: Best subset search. *Appl. Statist.*, **20**, 112–115.

Gentle, J. E. and Hanson, T. A. (1977) Variable selection under L_1. *Proc. Statist. Comput. Section, Amer. Statist. Assoc.*, 228–230.

Gentle, J. E. and Kennedy, W. J. (1978). Best subsets regression under the minimax criterion. *Comput. science & statist.*: *11th Annual Symposium on the Interface*. Inst. of Statist., N. Carolina State University, pp. 215–217.

Gentleman, W. M. (1973) Least squares computations by Givens transformations without square roots. *J. Inst. Maths. Applics.*, **12**, 329–336.

——— (1974) Algorithm AS 75. Basic procedures for large, sparse or weighted linear least squares problems. *Appl. Statist.*, **23**, 448–454.

Goodman, L. A. (1971) The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, **13**, 33–61.

Gunst, R. F. and Mason, R. L. (1980) *Regression Analysis and its Application*. New York: Marcel Dekker.

Hammarling, S. (1974) A note on modifications to the Givens plane rotation. *J. Inst. Maths. Applics.* **13**, 215–218.

Hemmerle, W. J. and Carey, M. B. (1983) Some properties of generalized ridge estimators. *Commun. in Statist.*, **B12**, 239–253.

Hjorth, U. (1982) Model selection and forward selection. *Scand. J. Statist.*, **9**, 95–105.

Hocking, R. R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

——— (1983) Developments in linear regression methodology: 1959-1982. *Technometrics*, **25**, 219–230.

Hocking, R. R. and Leslie, R. N. (1967) Selection of the best subset in regression analysis. *Technometrics*, **9**, 531–540.

Judge, G. G. and Bock, M. E. (1978) *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. Amsterdam: North Holland.

Kahaner, D., Tietjen, G. and Beckman, R. (1982) Gaussian-quadrature formulas for $\int_0^\infty e^{-x^2} g(x)\, dx$. *J. Statist. Comput. Simul.*, **15**, 155–160.

Kennedy, W. J. and Bancroft, T. A. (1971) Model building for prediction in regression based upon repeated significance Tests. *Ann. Math. Statist.*, **42**, 1273–1284.

Kohn, R. (1983) Consistent estimation of minimal subset dimension. *Econometrica*, **51**, 367–376.

Kudo, A. and Tarumi, T. (1974) An algorithm related to all possible regression and discriminant analysis. *J. Japan. Statist. Soc.*, **4**, 47–56.

Lawrence, M. B., Neumann, C. J. and Caso, E. L. (1975) Monte Carlo significance testing as applied to the development of statistical prediction of tropical cyclone motion. *4th Conf. on Prob. & Statist. in Atmos. Sci., Amer. Meteor. Soc.*, pp. 21–24.

McQuilkin, R. A. (1976) The necessity of independent testing of soil-site equations. *J. Soil Sci. Soc. of Amer.*, **40**, 783–785.

Miller, R. G. (1962) Statistical prediction by discriminant analysis. *Meteor. Monographs* (*Amer. Meteor. Soc.*), Vol. **4**, no. 25.

Narula, S. C. and Wellington, J. F. (1979) Selection of variables in linear regression using the sum of weighted absolute errors criterion. *Technometrics*, **21**, 299–306.

Nijenhuis, A. and Wilf, H. S. *Combinatorial Algorithms: for Computers and Calculators*. New York: Academic Press.

Pope, P. T. and Webster, J. T. (1972) The use of an *F*-statistic in stepwise regression procedures. *Technometrics*, **14**, 327–340.

Reingold, E. M., Nievergelt, J. and Deo, N. (1977) *Combinatorial Algorithms: Theory and Practice*. New Jersey: Prentice-Hall.

Rencher, A. C. and Pun, F. C. (1980) Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.
Roodman, G. (1974) A procedure for optimal stepwise MSAE regression analysis. *Operat. Res.*, **22**, 393–399.
Scheffé, H. (1959) *The Analysis of Variance*. New York: Wiley.
Spjøtvoll, E. (1972a) Multiple comparison of regression functions. *Ann. Math. Statist.*, **43**, 1076–1088.
——— (1972b) A note on a theorem of Forsythe and Golub. *SIAM J. Appl. Math.*, **23**, 307–311.
Steen, N. M., Byrne, G. D. and Gelbard, E. M. (1969) Gaussian quadratures for the integrals $\int_0^\infty \exp(-x^2) f(x)\,dx$ and $\int_0^b \exp(-x^2) f(x)\,dx$. *Math. of Comput.*, **23**, 661–671.
Tarone, R. E. (1976) Simultaneous confidence ellipsoids in the general linear model. *Technometrics*, **18**, 85–87.
Thompson, M. L. (1978) Selection of variables in multiple regression: Part I. A review and evaluation. Part II. .Chosen procedures, computations and examples. *Int. Statist. Rev.*, **46**, 1–19 and 129–146.
Wallace, T. D. (1977) Pretest estimation in regression: a survey. *Amer. J. Agric. Econ.*, **59**, 431–443.
Wellington, J. F. and Narula, S. C. (1981) Variable selection in multiple linear regression using the minimum sum of weighted absolute errors criterion. *Commun. in Statist.*, **B10**, 641–648.
Wilkinson, L. and Dallal, G. E. (1981) Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics*, **23**, 377–380.
Zirphile, J. (1975) Letter to the editor. *Technometrics*, **17**, 145.
Zurndorfer, E. A. and Glahn, H. R. (1977) Significance testing of regression equations developed by screening regression. *5th Conf. on Prob. & Statist. in Atmos. Sci., Amer. Meteor. Soc.*, pp. 95–100.

DISCUSSION OF DR MILLER'S PAPER

Professor J. B. Copas (University of Birmingham): I welcome Dr Miller to the Society, congratulate him on his presentation tonight, and thank him for bringing his paper out of his Private Bag into the public arena of one of our Ordinary Meetings. Dr Miller is surely right in saying that stepwise regression is one of the most widely used of statistical techniques. Thus tonight's review and analysis of the method is very much to be welcomed, and particularly so if, as I think it should, the paper helps to show that in many practical cases reliance on subset selection is misleading, wrong and foolish. Beloved of writers of statistical packages and users alike, subset selection is sadly lacking in a firm theoretical base. A discussion of the problems in this whole area is surely long overdue.

The most important aspect of tonight's paper is Dr Miller's repeated emphasis on these difficulties; simple selection methods fail to deliver, the usual significance tests are misleading, estimated regression coefficients are biased. Whilst agreeing with his emphasis on these difficulties, let me say why I think he should have gone further.

The lack of a firm theoretical base makes analysis of the properties of these methods almost impossible. I would welcome clarification from Dr Miller on what models are being assumed in the various parts of his paper. Surely the null hypothesis of zero regression coefficients for the omitted variables has itself depended on the data. How can we discuss estimation when the coefficient in question may or may not actually be estimated? In his likelihood method, Dr Miller conditions on the selected subset – but what justification can be given for this, bearing in mind that the choice of subset depends on the very same unknown parameters? More generally, I suggest that greater attention is needed to objectives. Are we assuming that the data are in fact generated by one particular subset, and that we are trying to discover which subset it is? Are we interested in which x's influence y? Is it prediction, and if so is it prediction at some given $\mathbf{x}$ or over some future population of $\mathbf{x}$'s? This last objective is by far the simplest, and some progress can be made as proposed in my own paper read to the Society last year (Copas, 1983). The earlier objectives, however, are quite a different matter. Required reading is Box's paper (Box, 1966), with its emphasis on the need for design when identifying the effects of individual regressors. No mention of design is made at all in tonight's paper, and I assume that most of the examples Dr Miller has in mind are observational in nature.

The root of many of the difficulties lies in the enormous scope for selection. Dr Miller takes the example of a search over $\binom{50}{5}$ subsets, saying that this large number, 2×10^6, is not beyond the powers of a modern computer. With somewhat fewer than 2×10^6 independent observations, perhaps 30 or 40, can we expect anything sensible at all to emerge? To translate this into a problem I can understand, suppose we have 2×10^6 observations from a normal distribution. We must not be surprised if a search reveals an observation 5 standard deviations from the mean. To explain such an observation as mere random error plays havoc with statistical intuition.

The most favourable model for subset selection is when the x's are orthogonal and when there is a clear distinction between one subset in which the p β's are all equal to β^*, say, and the complementary subset in which all β's are zero. Supposing that all the x's have mean 0 and variance 1, we have in the usual notation

$$\sqrt{n}\hat{\beta}_i/\sigma \sim N(\sqrt{n}\beta_i/\sigma, 1), \quad i = 1, \ldots, k.$$

Put $\beta = \sqrt{n}\beta^*/\sigma$. Then the probability that the selected subset of size p will be the correct one is

$$P(\min_{1,\ldots,p} |\hat{\beta}_i| > \max_{p+1,\ldots,k} |\hat{\beta}_i|) = p \int_{-\infty}^{\infty} \{\Phi(\beta - |t-\beta|) + \Phi(-\beta - |t-\beta|)\}^{p-1}$$

$$\{2\Phi(|t-\beta|) - 1\}^{k-p}\, \phi(t)\, dt,$$

where Φ and ϕ are the distribution and density functions of $N(0, 1)$ respectively. Note that the population multiple correlation coefficient is

$$R = \text{corr}\,(y, E(y \mid x)) = \sqrt{(p\beta^2/(p\beta^2 + n))}.$$

For example, taking $n = 100$, the graphs of Fig. D1 show this probability as a function of k and R at $p = 5$ and 10. When $R = 1$ the correct subset is always uncovered, but the probability falls away rapidly as R reduces from 1, and as k increases. For small R the situation is hopeless, with the selected subset almost certainly being wrong. Subset selection is perhaps used most frequently on social and medical data, for which $R = 0.5$ is certainly not unrealistically low, and in such cases it appears that almost no competition amongst x's can be tolerated. This analysis, of course, relates to a very special model, but practical situations are probably worse, with their wide spectra of valúes of the non-zero coefficients, with the search being over p as well as over subsets of size p, and with the added complexities of correlated x's.

Could research along these lines lead us to ban the use of subset selection altogether if R falls below some threshold defined in terms of p, k and n? Perhaps a warning message should be built

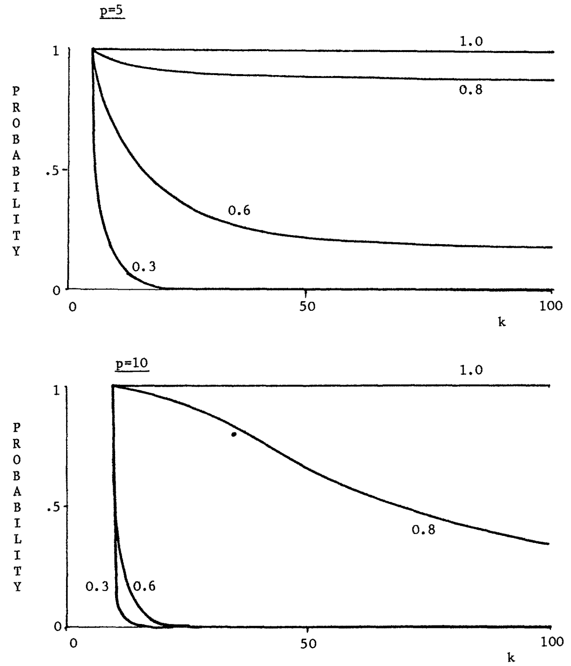

Fig. D1. The probability of selecting the correct subset (label on curve = value of R).

into all subset regression programmes. My guess is that, were this to be done, users would obtain the warning more often than they would obtain any actual answers. It has been said: "If you torture the data for long enough, in the end they will confess." Errors of grammar apart, what more brutal torture can there be than subset selection? The data will always confess, and the confession will usually be wrong.

Perhaps I am being unduly pessimistic. There are many counter examples in the literature which show how misleading subset selection can be, but where are the success stories? Dr Miller himself has estimated that something of the order of 10^5 multiple regressions are carried out per day worldwide, many of these using subset selection. If the method is of any value, there must surely be many interesting case studies which would prove me wrong. Where are they?

Dr Miller has challenged us tonight to consider a frequently used, a frequently abused and a little understood method of applied statistics. If his paper sparks off further research leading to a clearer understanding of subset selection and of when the method should and should not be used, it will have earned its place as an important contribution to the statistical literature. I have much pleasure in proposing the vote of thanks.

Professor M. Stone (University College, London): In 1924, R. A. Fisher, doing something about the weather at Rothamsted, wrote:

> "A still more insidious source of illusory high correlations lies in the fact that the particular varieties, chosen for correlation with the crop figures, are often chosen *because* they appear in fact to be associated with the crop."

Fisher then characteristically posed, and uncharacteristically failed to solve, a precise problem concerning this selection process, namely, the distribution of R for the best subset of size $q < n$ in the spherically random case. In the same decade, American psychometricians were addressing the same question and, by the 1940s, were proposing split-half assessment as a technique for encouraging realism in the selection of variables for test construction.

If only Fisher – or someone – had solved some even vaguely relevant, null problem and produced a set of significance tables, it is possible that the naive exploitation in this area of the computer power of recent decades would never have occurred.

I must confess I very nearly contributed – on the naivety side – by some picking and choosing of subsets of variables on an IBM 1620, generating the dubious inference of *negative* marginal utility of land to Irish farmers, suggesting it might be worth Irish farmers paying someone to steal some of their land. Fortunately, the paper was not accepted for publication. As it was, the need to do your own selection did allow some feelings of reservation about the output to surface.

The next generation of computers and packages, however, with their solemn and methodologically vacuous prescriptions, encouraged users to suppress any such reservations. It is this tide of misuse that Dr Miller's paper is so rightly concerned to confront.

My main comment on the paper is that it does not discuss the possibility that the best hope for dealing with the problem is not to wait for a theoretical solution of either Fisher's or any similar problems in mathematical statistics – which are always likely to be problem specific – but to arm the latest generation of computers with refinements of the psychometrician's split-half technique. And, where such refinements cannot usefully be implemented, perhaps to concede with Irwin Bross (1982) that "in practice the alternative to a simple analysis is not necessarily a more sophisticated analysis, it may well be no analysis at all".

I have some specific comments that would have been preferably put directly to the author, had I refereed the paper. So I will merely list them as bald assertions, and invite Dr Miller to accept or refute them in his last words on the discussion.

(i) The emphasis of the paper on least squares is not innocuous. The admission of other estimators can, at the cost of increased prior input, affect drastically the need for selection.

(ii) Condition (2.1) defies the sufficiency of "Pythagoras".

(iii) The "alternative method" of randomly generated dummy predictors has the zany logic of a radio monitor who tries to refresh his analysis of a weak, noisy signal by listening to some pure white noise for comparison.

(iv) In (5.2), $e(x_A)$ is undefined: $e(X_A)$ is a vector defined anew for each X_A.

(v) The Hjorth paper goes much further than is suggested by Dr Miller's dismissive comment.

Mention of the findings of Hjorth brings me full circle – to the cross-validators of the 1930s! Additional evidence of the value of that approach in controlling – without assumptions – the excesses of any proposed selection process is to be found in Mabbett *et al.* (1980) in the area of medical diagnosis and in O'Brien *et al.* (1984) in the context of radio-activation analysis. The book of Breiman *et al.* (1983), which I have not yet seen in its final form, promises a general framework for interesting applications.

Dr Miller has been thoroughly provocative in his review of what may be the most pressing problem area in Statistics—and I am happy to second the vote of thanks.

The vote of thanks was carried by acclamation.

Dr R. L. Plackett (Retired): Dr Miller's critical examination of this field is very welcome, in view of the many methods that have been proposed. I have learned much from his paper, and from a selected subset of the references.

His assumption (1.1) raises the question of how exactly the deterministic error $e(X)$ is defined, and the comments in different sections point in different directions. Thus Section 1 refers to the best prediction in some sense, averaged over a region, and suggests that the underlying model is a least squares regression function derived from a general relationship, whereas Section 5 indicates that $e(X)$ arises from non-linearity or the omission of interactions. The matter is rather puzzling, and I would welcome clarification of what it is that the subsets fit well if the vector of residuals is zero.

I would like to turn next to the discussion of bias, which Dr Miller has illuminated with careful analysis and stimulating ideas. Method (v) in Section 5 looks formidable, but even when the difficulties are overcome there is no reason why maximum likelihood should give unbiased estimates of the regression coefficients when the model is selected by processing the data. However, there are various reasons for thinking that the effects of bias can be exaggerated. The illustration in Section 5 is based on a sample of size 31 which for most purposes would be considered small. Even so, inspection of Fig. 2 suggests that both true and false MSEP leads to the same conclusion if we agree to stop when further reductions are either small or cease altogether. Those of us who toil in the fields of categorized data have long been accustomed to biased estimators and tests that are valid only asymptotically. Methods which distinguish between the distribution of F and χ^2 are more refined but less robust. Models for contingency tables always include the term $e(X)$ that makes fleeting appearances in the paper, and the selection is usually made from the class of hierarchical models. The number of such models is unknown except for tables of small dimensionality.

Notwithstanding the many interesting results given here, I disagree with Dr Miller's concluding comment that further research is needed on the reduction of bias, and not only for the reasons given. If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems.

Professor M. A. Aitkin (University of Lancaster): It is a pleasure to welcome Dr Miller to tonight's meeting, and to be able to comment on his interesting paper. Dr Miller notes that there are several different reasons for using variable subsets. The main distinction is between description and prediction.

For description the aim is parsimony. Different subsets of variables may provide adequate representations of the data, especially in small samples with correlated explanatory variables. In such cases it is important to present the different candidate subsets and not to draw strong conclusions.

The determination of adequate subsets through a simultaneous test procedure was given in Aitkin (1974). Here all subsets are being compared with the full model, not with each other as in the Spjøtvoll test, because if two different subsets are both adequate representations of the data, their direct comparison is of no interest. The set of adequate models can be characterized by the *minimal* adequate subsets, which themselves have no proper adequate subsets.

In the STEAM data, for example, there are 9 predictors and 25 observations. For the full model the residual sum of squares is 4.87 on 15 d.f. with $R^2 = 0.924$. A simultaneous test of size α rejects any model as adequate if its residual sum of squares exceeds $(1 + 9F^{\alpha}_{9,15})$ 4.87. Taking $\alpha \cong 1 - (0.95)^9 = 0.369$, or $\alpha = 0.35$, the limit is 8.45, corresponding to $R^2 = 0.868$. None of the two-variable sets in Table 2 provides an adequate description at this level (the best is just adequate if $\alpha = 0.275$). The subset comparisons in Table 4 are thus of little interest: either both subsets A and B are adequate, in which case we cannot choose between them except on the grounds of parsimony, or A and/or B is not, in which case the inadequate ones would not be used.

For prediction, a Spjøtvoll-type test is valuable because we *do* want to compare different subsets to find those with small *MSPE*. The *STP* in Aitkin (1974) only compares subsets with the full model, which is not sufficient. In evaluating the *MSPE*, it seems preferable to condition on the new **x** rather than to average over it. This should give smaller *MSPE* for each new observation, at the expense of more computation, since the subset to be used for prediction may depend on the new **x**, as Dr Miller noted in his presentation.

Professor D. V. Lindley (Somerset): There are many who think that to have a coherent philosophy about a subject is a luxury for cloistered academics and that operators can get by with strictly pragmatic considerations. Tonight's paper is a good counter-example because the lack of a coherent view, and the failure to cite references to that view, leads to many unsatisfactory arguments. Here are three examples.

The paper uses least squares: yet this method has been known for 27 years to be inadmissible. Its efficiency can be as low as $2/k$ for k predictors.

As a consequence of using least squares, troubles arise in selecting a subset. Breiman and Freedman (1983) show that least squares leads to a best subset and that the inclusion of variables outside that set will increase the error of prediction. Let $\mathscr{F}_k$ be the class of linear predictors using k variables. Clearly $\mathscr{F}_{k-1}$ is a proper subclass of $\mathscr{F}_k$ obtained by putting the coefficient of X_k in the latter zero. Consequently any function minimized over $\mathscr{F}_{k-1}$ cannot reach a lower value than over $\mathscr{F}_k$ and prediction using $k-1$ cannot be better than with k. The fallacy arises because least squares gets worse as k increases. The last argument shows that the more variables the better. Selection can only be justified on utility or loss considerations which are entirely lacking from the paper.

Thirdly, consider the question of bias in Section 4. If data D have been used to select X_1 and discard X_2 the prediction of Y given X_1 coherently requires the calculation of $p(Y \mid X_1, D)$. This is

$$\iiint p(Y \mid X_1, X_2, \beta_1, \beta_2, D)\, p(X_2, \beta_1, \beta_2 \mid X_1, D)\, dX_2\, d\beta_1\, d\beta_2$$

$$= \iiint p(Y \mid X_1, X_2, \beta_1, \beta_2)\, p(X_2 \mid X_1, D)\, p(\beta_1, \beta_2 \mid D)\, dX_2\, d\beta_1\, d\beta_2$$

under reasonable assumptions (a) of exchangeability between the prediction set and D, and (b) that X_1, X_2 give no information about β_1, β_2. This tells us what to do. Clearly it is not equivalent to the prediction $b_1 X_1$. Omission bias is allowed for: competition and stopping-rule biases do not enter. One of the more difficult problems in life is to recognize what questions are sensible to ask. Attempts to answer unsatisfactory questions are bound to be unsatisfactory. Questions of bias are exactly of this sort and do not arise in a coherent view.

Whilst the emphasis in this paper is wrong, the author has thought a lot about the problems in this difficult field and the paper contains many insights that are valuable in any understanding of it. Some of the bricks are excellent: the cement is a bit cracked or even lacking in places.

Professor E. M. L. Beale (Scicon Ltd): Dr Miller's valuable paper reminds us that, even though multiple regression is very useful for summarizing data and suggesting models, it may be misleading if given a precise statistical interpretation.

Dr Miller criticizes Gauss–Jordan methods for selecting subsets as numerically inaccurate. But they need not be. We all now know that we must not select a variable that is highly correlated with those already selected. But we must also not select a variable that would make any previously selected variable highly correlated with the others. Frane (1977), and more explicitly Clarke (1982), show how to test for this when using Gauss–Jordan methods. With *QR* methods, subsequent calculations are not affected by the inadvertent selection of a linearly dependent subset, but the test should still be made and is much more laborious.

Dr Miller refers to Furnival and Wilson (1974), who changed the optimum subsets algorithm of Beale *et al.* (1967) in three main ways. One is to select the best 10 or so equations with each number of variables. Another is to keep $k+1$ versions of the partially inverted correlation matrix, rather than just 2, to avoid repeating steps. These I thoroughly accept. The paper does not explicitly discuss linear dependencies among the independent variables: when detected, these make the algorithm more cumbersome but are not disastrous. The other change is to omit the row and column for a variable once it has been selected. This I do not like, because it prevents both the use of the Frane check and the rejection of physically meaningless negative regression coefficients.

When $k \gg 20$, one should be able to limit the number of alternative equations explored with each number of variables, and to accept the best solutions found so far: such facilities are widely used in the related discipline of integer programming as a way of controlling the computing cost.

Section 5 discusses mean square prediction errors when non-significant regression coefficients are set to zero. These can easily be computed as functions of the true regression coefficients when σ^2 is known and X^TX is diagonal. They are maximized when all the true values are near their significance levels. Sprevak (1976) studied the bivariate case when X^TX is not diagonal, and showed that the maximum mean square error is not greatly increased. This is encouraging, although one would like to see the work extended to higher dimensions.

Dr I. T. Jolliffe (University of Kent at Canterbury): The idea, mentioned in Section 2, of breaking the predictors into groups and performing searches only *within* each group is an interesting one. A naive approach would be to consider including only one variable from each group, provided that all variables within a group are highly correlated.

A frustrating aspect of variable selection in regression or elsewhere is that for any potential method, except those involving an exhaustive search of all subsets, it seems possible to construct simple examples for which the method does not work. The artificial data set in Table 1 of the paper is a good example; not only does it demonstrate the points discussed in Section 2, but it also shows that the naive approach outlined above, is flawed, the crucial feature of the example being that X_1, X_2 are highly correlated and are therefore in the same group.

It is not, however, necessary for X_1, X_2 to have a large correlation for the other properties of the example to hold. It is possible to construct a simple example for which $Y = X_1 - X_2$ and X_1, X_2 are uncorrelated, but another variable X_3 is picked first, with an insignificant reduction in residual sums of squares when either X_1 or X_2 is added to X_3. Such behaviour should not cause problems in small examples where an exhaustive search is possible, but they illustrate serious potential difficulties if X_1, X_2, X_3 are part of a much larger set of variables.

Another suggestion in Section 2 is to add, delete or replace variables two at a time. In a recent PhD thesis, Dr R. E. Kempson of Wye College has shown that forward selection of variables in discriminant analysis can, in some circumstances, be substantially improved, with relatively little additional computing cost, by adding variables two or three at a time rather than the usual one-at-a-time. However, deletion of more than one variable at a time in backward elimination was generally less effective.

Dr R. W. Farebrother (University of Manchester): Dr Miller's use of p differs from that of Mallows (1973) so that his definition of C_p should be formally increased by two. With this modification $s^2 C_p = \text{RSS}_p + (2p + 2 - n)s^2$ is an unbiased estimator of $E\,\|X\beta - X_A b_A\|^2$ provided that $X = [X_A\ X_B]$ has been partitioned independently of y. However, if we choose the partition which minimizes C_p then the optimal partition is a function of y and the use of the criterion destroys the foundations on which it was erected.

This criticism may be lodged against all estimators of the form $\tilde{\beta} = D_*'X'y$ where D_* is chosen to minimize

$$M_1 = (\tilde{\beta} - b)'\,W(\tilde{\beta} - b) + s^2 \operatorname{tr} WD'X'XD$$

or

$$M_2 = (\tilde{\beta} - b)'\,W(\tilde{\beta} - b) + s^2 \operatorname{tr} W[2D - (X'X)^{-1}]$$

subject to constraints on the form of D and where M_1 is a biased estimator and M_2 an unbiased estimator of

$$M_0 = E(\tilde{\beta} - \beta)'\,W(\tilde{\beta} - \beta).$$

If $W = X'X$ and $D = \operatorname{diag}\{(X_A'X_A)^{-1},\ O_{k-p}\}$ then $\tilde{\beta}$ is the subset estimator and $M_2 = s^2 C_p$, see Farebrother (1980).

The subset estimator is also the limiting member ($h \to \infty$) of the class of partitioned ridge estimators

$$\begin{bmatrix} \tilde{\beta}_A(h) \\ \\ \tilde{\beta}_B(h) \end{bmatrix} = \begin{bmatrix} X_A'X_A & X_A'X_B \\ \\ X_B'X_A & X_B'X_B + hI_{k-p} \end{bmatrix}^{-1} \begin{bmatrix} X_A'y \\ \\ X_B'y \end{bmatrix}.$$

Dr J. R. Green (University of Liverpool): Congratulations are due for this excellent exposition of the "state of the art" covered by the title of this paper, with its masterly examination of the rival methods that are – or may be used in this area.

I wish to mention a further "cheap method" that appears in a related paper by Mustaffer Al-Bayatti and myself which has been submitted .to *Mathematische Operationsforschung und Statistik*, Series Statistics, entitled, "Selection of regressor variables when $E(Y)$ is an unknown nonlinear function". This paper approximates to the unknown hypersurface of EY in terms of the x's by a set of hyperplanes, and then the Beale-Kendall-Mann procedure for selection of regressors is employed in that situation. Some criticisms of many methods used apply to our paper too. However, as in the ordinary linear situation (only more so), there are sometimes too many regressors to handle, so a screening method is suggested. Here those regressors are removed for which the increase in the sum of squared residuals, when each is removed from the full set of regressors one at a time, is sufficiently small. The residual sum of squares for the remaining subset, after removing *all* those rejected regressors, is then compared with the RSS of the full regression as a check.

Dr R. F. Gunst (Southern Methodist University, Dallas, USA): My remarks address only one aspect of Dr Miller's discussion of variable selection, the influence of collinearities on competition for selection. Collinear predictor variables necessarily compete for inclusion in "best" subset algorithms because by their very nature they are redundant, at least in the data set being analysed. Collinearities can accentuate the effects of competition for selection. For instance, 9 of the 11 predictors in the Detroit homicide data have at least one strong pairwise collinearity with $|r| > 0.90$. The biases evident in Table 5 neither reflect correlations as large as this nor do they account for multiple collinearities.

No variable selection technique has demonstrated the ability to consistently identify the "best" subsets in repeated sampling when predictor variable are highly collinear. Least squares in particular has been shown to be ineffective. Likewise the several examples in Section 3 illustrate the failure of least squares algorithms to produce consistent results on a single data set. The replacement algorithms suggested in this article are major improvements over one-at-a-time selection algorithms, but they too have pragmatic limitations. Not only are r-variate replacement algorithms difficult to code, they suffer from defects similar in nature to those of one-variate selection routines; e.g. r-variate replacement algorithms can fail to detect $(r + 1)$-variate synergistic effects.

The identification of collinearities not only alerts the analyst to the presence of collinearities but it also suggests tradeoffs which most variable selection techniques do not identify. For example, if the pairwise correlation between two variates is, say, as large as 0.99, either variate should be a candidate for selection regardless of the preference of variable selection procedures (provided that at least one of the two is selected). Note too that identification of collinear variates can reduce the amount of variable selection that need be performed because there generally is consistency in the selection of influential non-collinear variates with most of the variable selection techniques discussed in this article.

A final caution about collinear effects on competition for selection concerns the fitting of polynomial regression models. Severe collinearities can be induced and competition bias can be overwhelming when polynomial or interaction terms are added to linear terms in a regression model. A good example is provided by the nitrous oxide data set in Gunst and Mason (1980, p. 362).

Professor D. A. Harville (Iowa State University, Ames, USA): Dr Miller's paper provides many valuable insights into the problem of selecting regression variables. Research on this problem seems to have been plagued by confusion as to the nature of the objectives. My discussion concerns these objectives and their implications.

In the subset-selection problem, we use the data to choose a subset S from k regression variables $x_1, \ldots, x_k$ and then, for example, in predicting the realization of a random variable y, adopt a predictor of the general form $\hat{y} = b_0 + \sum_{j \in S} b_j x_j$. Corresponding to the predictor $\hat{y}$ is an estimator $\hat{\beta}_j$ of β_j defined by $\hat{\beta}_j = b_j$, if $j \in S$, and $\hat{\beta}_j = 0$, if $j \notin S$.

In his Section 1, Dr Miller lists possible reasons, numbered (i)-(iv), for using only a subset of the regression variables. It would seem that the procedure for choosing S and the formulas for the coefficients b_0, b_j $(j \in S)$ should be determined on the basis of a criterion that is consistent with the relevant reasons. Yet, such an approach seems seldom to be taken.

It is evident that unless the criterion includes a penalty for choosing S to be the set $\{1, \ldots, k\}$ of all regression variables, it will not distinguish, to any meaningful extent, between two predictors like $1 + x_1$ and $1 + x_1 + 10^{-30}(x_2 + \ldots + x_k)$. When Dr Miller's reason (i) is relevant, the criterion should reflect the lower cost that results from observing a smaller number of regression variables. For example, following Lindley (1968), we could proceed on the basis of the loss function $(\hat{y} - y)^2 + c_S$, where c_S is a real cost associated with observing the values of x_j $(j \in S)$. The same criterion could be used in conjunction with Dr Miller's reason (iii), except now c_S would represent an artificial cost that reflects non-parsimony, rather than a real cost.

A setting in which Dr Miller's reason (ii) or (iv) is relevant would seem to call for a criterion, like (true) *MSE*, that does not include an explicit penalty for large subsets. As is evident from Dr Miller's simulation study of the PLANES example, it is important to account for selection biases when evaluating the *MSE* of any predictor or estimator that incorporates subset selection.

For a satisfactory resolution of the subset-selection problem, we can look to the class of (proper and improper) Bayes procedures. The Bayes approach was considered by Lindley (1968), in a paper which I regard as the most important yet to appear on subset selection. A frequentist can choose from the class of Bayes procedures on the basis of their frequentist properties (or can, by employing an empirical Bayes approach, allow the data to make the choice).

Professor M. J. R. Healy (London School of Hygiene): The accepted wisdom on the selection of regression variables has changed over the years in an interesting way. Early on, the use of R (or equivalently, the residual sum of squares) as a criterion suggested that adding a variable always improved the situation. Very soon it was realized that a more realistic criterion was the residual mean square (or R adjusted by shrinkage); now adding a variable might make things better but could not (on average) be detrimental. The next step was the use of C_p and *AIC* which indicated that more could actually mean worse. Now Dr Miller and Professor Copas both appear to tell us that we should expect prediction to go on improving as variables are added and that the apparent advantages of parsimony are illusory. If I understand Professor Lindley, he holds this result to be self-evident on logical grounds.

I think two qualifications are needed to these findings. Dr Miller's Fig. 2, which shows a steadily decreasing *MSEP* as x's are added, is based on empirical data. In such data it is highly unlikely that any of the x's is *completely* unrelated to the predictand, and Professor Lindley's argument holds. Variable selection methods using t or F criteria are really aiming to eliminate x-variables whose true coefficients are *exactly* zero (just as do the orthodox significance tests from which they arise). The implausibility of the occurrence of such variables in practical situations is another argument against the appropriateness of these methods.

If in practice more means better, if logically it cannot mean worse, where is the argument for parsimony, needed as it is to back up one's intuition that going on adding x's indefinitely cannot be a very sensible thing to do? I think the argument must be based on a realistic appreciation of the fact that providing the values of x-variables and using them for prediction and control are not cost-free activities. A parsimonious regression formula may pay for a suboptimal level of prediction or control by economies in data collection and manipulation. In the absence of some such approach, I agree with Dr Miller that much of a massive literature is of little relevance to practical data analysis.

Dr Urban Hjorth (Linköping Institute of Technology): I congratulate Alan Miller on a very interesting paper about model selection in regression. Some years ago one branch of model selection seemed to have come to an end with the development of Akaike's AIC criterion and some consistent refinements. To me some regression problems in connection with meteorological data, with several predictors involved, clearly demonstrated the very strong bias of the usual estimates of mean square error of prediction (MSEP). Like Miller we also found that AIC and related criteria did not give relevant correction of this bias. For this estimation problem we have good experiences of using cross validation and forward validation (Hjorth and Holmqvist, 1981; Hjorth, 1982). Miller addresses in Section 4 a different problem, namely the bias in the regression coefficients themselves and mentions five approaches to remove this bias. I appreciate Miller's clear distinction between various sources of bias (due to omission, competition and stopping) and I quite agree with his statement that competition bias is very important and far more serious than stopping rule bias in most situations of interest.

An interesting question is the effect of the five methods in Section 4 on the MSEP. Can data split into halves ever be efficient in this respect? Jack-knife, bootstrap and the Monte Carlo approach (iv) all seem to suffer from the complication that the selected set of predictors will vary. If many predictors are involved a quite small proportion of the results may be useful if we "condition" on the same selected set of predictors. That shrinking can reduce the MSEP was recently demonstrated to the Society by Copas (1983). The conditional likelihood approach is interesting. Can we expect more from this method than from shrunken estimators in terms of MSEP, and can this method handle more than small problems? Miller's paper poses several interesting questions and will no doubt stimulate further research about subset selection and estimation methods. I want to thank him for this opportunity to read and comment upon his excellent paper.

Dr C. L. Mallows (AT & T Bell Laboratories, NJ, USA): It is somewhat disturbing to have one's work both unreferenced and dismissed as invalid. In my 1973 paper I derived the C_p statistic (scaled by σ_A^2) as an unbiased estimate of the MSEP (except for the σ_A^2 term), when the subset of variables used is held fixed, and bias terms (such as those in Miller's (5.2)) are allowed for. No claim was made that the minimum realized C_p was an estimate of the MSEP of any (selection+prediction) rule; in fact in my Section 4 I went to some pains to point out that this quantity, using the minimum-C_p rule for selection, could differ considerably from the minimum C_p estimand. I agree that it is disappointing that we do not know how to estimate the MSEP, when the selection effect is allowed for, but I would give higher priority to finding a good selection rule than to finding a good estimate of the expected performance of a possibly inferior rule.

Absolute priorities are another matter. When the situation is ambiguous, and inspection of the C_p plot will show this, no (selection+l.s.prediction) rule can do well. Also, the context of the problem must enter into the choice of a predictor.

I also gave (in Section 3 of my paper) a "valid" calibration of the C_p plot, using standard hypothesis-testing methods.

I do not see the point of distinguishing three sources of bias. All can be ascribed to inappropriate selection.

Dr C. A. Platt (San Francisco State University): The author is to be congratulated for providing a most provocative and stimulating paper. As he indicates, estimation bias results from using the same data set for both selection of variables and estimation of parameters. This bias has been decomposed into omission bias and selection bias, with the latter further decomposed into competition bias and stopping-rule bias. The author alludes to but does not develop the potential usefulness of the bootstrap resampling plan in reducing both types of selection bias.

The bootstrap substitutes computing power for statistical theory, thereby obviating the need to make distributional assumptions in assessing properties of a population from a sample data set. This is accomplished through simulations in which uniform random drawings with replacement from the sample data set serve as proxies for independent samples drawn from the underlying population. Statistics associated with these multiple samplings form a basis for inferences concerning population parameters. The only necessary assumption is that the simulated replicates bear the same resemblance to the distribution of sample data from which it is drawn that the sample (and other samples, if available) bears to the underlying population.

The bootstrap can be applied several ways in the present regression context. If it is used only for variable selection or only for parameter estimation, or if it is used for both but separate resampling plans are employed for each, bias from using the same set of data for both purposes will be reduced. It could be applied to estimate the true but unknown number of variables which properly should enter the model, possibly through examination of bootstrap replicates from a "cheap" method such as the Efroymson forward stepwise algorithm. Knowing this parameter, it usually is feasible to examine all possible subsets of that specified size. If either problem size or computing constraints prevent an examination of all possible subsets, and if branch-and-bound algorithms are not of sufficient assistance in the selection of variables, the bootstrap could again be employed. One possible strategy would be to examine the variable selection scheme from stepwise regressions based upon bootstrap resamplings of the data. Among those runs in which the previously estimated correct number of variables entered, identify that variable which entered most frequently, then that variable which entered most frequently those models that also include the first variable, and so on until the specified number of variables have been selected.

Dr D. A. Preece (Rothamsted Experimental Station): Just before the meeting, Dr Miller said to me that his earlier estimate of 100 000 multiple regression analyses per day was probably an *under*estimate—perhaps even by a factor of ten. He would make no estimate, however, of the proportion of *sensible and worthwhile* analyses! Whether that proportion be one in a hundred or one in a hundred thousand, there must still be a lot of multiple regressions around that merit statisticians' critical attention—and this leads to a difficult question: How could a set of examples be compiled that might in some sense be "representative" of all the multiple regressions that thinking scientists and economists wish to do? Such a set is desirable both for illustrating papers such as Dr Miller's and for teaching.

The monthly STEAM data, for instance, have several features to distinguish them from other examples. These data are clearly for twenty-five consecutive months, January to January. Obvious preliminary questions are: (i) Might it be better to use only a "balanced" set of twenty-*four* consecutive months?, and (ii) Ought there be at least one x-variate to take account of the time-sequence? Dr Miller's predictors 1, 2, . . ., 9 are (I think) Draper and Smith's $X_2, X_3, \ldots, X_{10}$ respectively. .Of these, X_9 is (apart from some dotty rounding) the square of X_4; bringing X_9 into a model before X_4 (Dr Miller's Table 4) therefore needs comment, at the least. Of the eight predictors excluding X_9, four happen to be discrete, each with few distinct values; these four are X_5 (no. of calendar days), X_6 (no. of operating days), X_7 (no. of days with freezing temperatures) and X_{10} (no. of "startups"). The main feature of X_6 is the annual holiday for about half of each July; variate X_7 too is very skew, with zero values throughout each summer; variate X_{10} has values 2, 3, 4, 5 and 6 only. Some comment on all this is needed, I believe, if the data are to be used for an "illustrative" example: data with four such discrete variates and four continuous ones may be "representative" of something, but of what? Also, knowledge of what the variates are, makes it unlikely that the y-variate X_1 (pounds of steam used) truly depends on X_5, and very unlikely indeed that X_1 has any dependence on X_5 that is not covered by dependence on X_6 (again cf. Dr Miller's Table 4); to do multiple regression analyses and discuss them without naming the variates and taking account of the names is at the very least a risky "illustrative" exercise.

The **Author** replied later, in writing, as follows.

I would like to thank all of those who contributed to the discussion, including some with whom I do not completely agree. Subset selection in regression has often been described colourfully as, for example, fishing expeditions, torturing the data until it confesses, or data dredging, so I did not expect to be treated quite as politely! However, as so much of this data dredging is being done, I feel that it is time that someone documented its problems and limitations, and the practice is either abandoned or put on a sound theoretical footing. I do not claim to have achieved this, but I hope that my paper and this discussion at least disturbs a few of the practitioners and stimulates research in this field.

In my oral presentation, I said that my paper was vulnerable as I was not addressing one specific objective of subset selection in model building, and was trying to provide something for all users while satisfying none. Several discussants have chosen to speak on the subject of objectives, particularly Professor Harville who has partially quantified the four objectives which I listed.In any particular application, the choice of a subset must depend upon what is to be done with the model, and in some situations two or more subsets should be chosen for different objectives even though the same calibration data are being used in each case. For instance, the work of Galpin and Hawkins (1982) suggests that, when the objective is prediction with no cost for the measurement of variables, it will sometimes be desirable to use different subsets in different regions of the (total) X-space. My own work shows that a quite different stopping rule will usually be necessary if least-squares estimates of regression coefficients are used rather than estimates which make an allowance for selection bias.

Unfortunately, the PLANES data were not published in Copas (1983) though the data were in the draft which I received.

Both Professor Stone and Professor Plackett asked about the definition of $e(X)$, which first appears at the end of Section 1. If the true relationship is

$$Y = \eta(X) + \epsilon,$$

and we approximate the regression function, $\eta(X)$, by $X\beta$, then β could be defined as

$$\beta = (X'X)^{-1}X'\eta(X),$$

leaving the deterministic error as $\eta(X) - X\beta$.

Professors Copas and Healy discuss the hypothesis that $\beta_i = 0$ for the variables which are not selected. In most practical cases it will be unrealistic to assume no relationship at all for the omitted variables, and I prefer the approach of Spjøtvoll of placing confidence limits on the contribution of these variables to the regression sum of squares, rather than on the significance-testing approach. While Professor Copas's derivation of the probability of correct selection is interesting, I would think that it is rarely sensible to think in terms of a correct model; indeed that in itself is a contradiction of what constitutes a model. There is some literature on the elimination of poor subsets in such a way that the probability of eliminating the "true" or "best" subset is guaranteed to be small. This is often only a feasible approach when the number of available predictors is fairly small. The most recent reference which I have in this area is Huang and Panchapakesan (1982), which contains references to earlier work. Spjøtvoll's method appears to provide a good alternative method for eliminating the inferior subsets.

In expressing the hypothesis in the form

$$H\colon \beta_i = 0 \text{ for all variables not selected by procedure } XYZ,$$

I was attempting to say what hypothesis I thought the users of F-to-enter's and such-like tests are attempting to test. There is no conceptual difficulty in using this form of hypothesis, provided that the test statistic and procedure XYZ are precisely specified, though it may well be almost impossible to derive the distribution of the statistic under the null hypothesis in most cases.

I appreciate Professor Copas's difficulties in accepting my likelihood method. The decision to choose a particular subset has no place in likelihood theory, and yet I am trying to bend the likelihood idea so that it compensates for a decision which is probably wrong. The only crude justification that I can give is that if L is the log-likelihood defined in (4.7), then $E(\partial L/\partial\theta) = 0$ for any parameter θ, where the expectation is taken using the conditional density over the region R. The fact that this measure of central tendency is unbiased does not mean that the estimates are unbiased, but it does give grounds for being optimistic that the bias resulting from maximizing (4.7) may be small. This likelihood, if it can be called that, appears to be very well behaved. It appears to have only one maximum, despite attempts on my part to construct cases with a second local maximum; it has no discontinuities and all of its derivatives exist. My limited practical experience to date suggests that it reduces the selection bias by about 60 per cent.

A practical solution to the problem of finding the distribution of R, or rather of R^2 for the spherically random case mentioned by Professor Stone, has been given by Rencher and Pun (1980), as mentioned in the paper. I think I have convinced the authors of one statistical package to incorporate this test in its stepwise procedure. The Rencher and Pun formula is for the Efroymson algorithm and with predictors correlated in a particular way. It is a relatively straightforward exercise to obtain similar formulae for other cases. For instance, a colleague recently wanted to know the distribution of R^2 under the null hypothesis when forward selection with an F-to-enter of 1.0 and the condition that the sample regression coefficients must have the "right" sign were used. I suspect that the non-significance of values of R^2 could kill off many of the stepwise regressions which must be the basic output of most statistical packages.

I apologize to Dr Hjorth if my brief reference to his work is dismissive as Professor Stone says. In the antipodes we see journals several months after the rest of the world, and I had not read his paper (Hjorth, 1982) at the time I submitted the first draft of this paper. Hjorth's paper illustrates the extent to which prediction errors can be underestimated using formulae for mean squared errors of prediction which make no allowance for selection, and also looks at the problems in applying cross-validation techniques. A report by Hjorth (1983) extends his results to model selection in general, e.g. in choosing a distribution to fit to a data set.

The method of splitting the calibration data set into random halves is obviously inefficient, as Dr Hjorth comments. I too have worked with meteorological data in recent years, and this is one of the few fields in which there are often sufficient data for splitting to be reasonable. It has been very educational to use one half of the data to select a subset, to calculate the least-squares estimates of regression coefficients and the usual estimated standard errors (which assume the model was chosen *a priori*), and then to compare these with the regression coefficients estimated for the same subset of variables but using the other half of the data. As the standard errors are underestimated, the shrinkage of the regression coefficients has often been by 6 or more of these standard errors even though the sample sizes were in hundreds.

The "zany" method, as Professor Stone describes it, of adding extra artificial variables is fairly widely used in my experience, and hence its inclusion in the paper even though I do not know of a reference in the scientific literature.

It is difficult to believe that the condition (2.1) can be new, but so far nobody has told me of its prior discovery. Let me further "defy the sufficiency of Pythagoras" with a simple geometric explanation. In Fig. D2, Y_p is the projection of Y upon the plane of two non-orthogonal predictors X_1 and X_2. The point O is the origin so that the regression sum of squares is the square of the length of OY_p. If we consider the regression upon X_2 first, then the projection of Y upon its direction is the length of OB. Alternatively, if we regress upon X_1 first (projection is OA), then the projection upon X_2 is AY_p which has the same length as OB. Thus the regression sum of squares for X_2 is the same whether it or X_1 is the first variable entered. As the correlations between variables equal the cosines of the angles between their directions, the result (2.1) can easily be verified. The method of planar rotations applied to the Cholesky factorization to change the order of variables in a regression, provides a third way of deriving the result. I may have defied the necessity of Pythagoras but not its sufficiency.

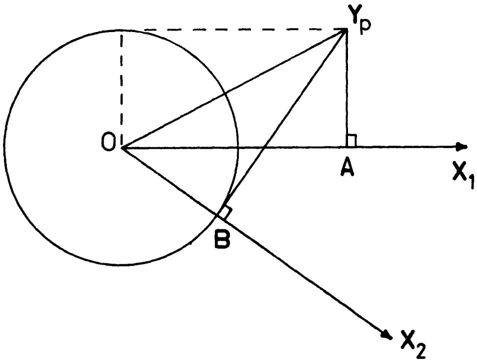

Fig. D2. Illustration showing the alternative case in which regression sums of squares are additive but in which the predictors are not orthogonal.

I welcome Professor Aitkin's contribution to the discussion. I believe that his adequate subsets method is identical with Spjøtvoll's method where the two are both applicable, though his is much simpler to apply. His use of a 35 per cent level of significance leaves me rather baffled. The arithmetic he uses to arrive at this figure, or actually at the 0.369, is of the kind which is used for independent tests, not for simultaneous tests.

The efficiency of $2/k$ which Professor Lindley quotes is for the estimation of regression coefficients, not for predictions. When Dr Farebrother was in Melbourne a few years ago, he demonstrated very effectively that while huge reductions in mean squared errors of regression coefficients are sometimes possible using shrunken estimators, the same estimators often give marginal reductions in mean squared errors of prediction. The family of estimators suggested by Dr Farebrother, i.e. $\tilde{\beta} = D_*'X'y$ seems attractive, once we can adequately approximate or bound the influence of selection on the objective function, such as his M_2. Perhaps an estimator can be found which gives good mean squared error performance for selection biases in the least-squares regression coefficients in the range 0–2 standard errors, where the bias is in the direction which makes the regression coefficient too large in absolute value. The following argument leads to this form of shrinkage. Consider the orthogonal reduction, $X = QR$, where $Q'Q = I$ and R is upper triangular. The projections of Y upon the orthogonal columns of Q are then the elements of $Q'Y$. Let us suppose that a subset A of p variables has been selected on some criterion, and that the columns of X have been ordered so that the first p columns correspond to those of subset A. The first p projections in $Q'Y$ will be biased on the high side, with the extent of the bias depending upon the amount of competition for selection. The least-squares estimates of the regression coefficients for this subset are given by

$$\hat{\beta}_A = (R_A^{-1}, O)\, Q'Y.$$

If we shrink each projection by varying amounts by multiplying by a diagonal (or non-diagonal) matrix D, then we obtain

$$\begin{aligned}\tilde{\beta}_A &= (R_A^{-1}, O)\, DQ'Y \\ &= (R_A^{-1}, O)\, DR^{-T}X'Y.\end{aligned}$$

The amount of shrinkage applied to each projection is then a combination of the shrinkage required to overcome the selection bias, plus that required to reduce the mean squared error of either the regression coefficients or the predictions, depending upon the application. Unfortunately the diagonal form of D will not generally be preserved when the order of variables in A is changed, and the diagonal form is perhaps appropriate only in the case of forward selection with the variables in the order of selection.

Professor Lindley's main point is that competition bias and stopping-rule bias do not enter in his derivation of his $p(Y \mid X_1, D)$. This is far from obvious, at least to me. It appears that he would use the same methods whether X_1 had been selected independently of the data or not. Suppose we divide the set of all possible sets of data (let us suppose the sample size is fixed) into two sets D_1 and D_2, such that if $D \in D_1$ then X_1 is the selected variable, otherwise X_2 is selected. Now using my pragmatic operator's approach to estimate (β_1, β_2), if I ignore the selection process and use a likelihood which does not integrate to one over D_1 or D_2, whichever was selected, then my β_1 is biased on the high side and β_2 on the low side in D_1 and vice versa in D_2. Alternatively I can use the kind of conditional likelihood which I use in Section 4 of the paper. Similar scope appears to exist for Professor Lindley in deciding what to do in estimating his $p(\beta_1, \beta_2 \mid D)$. Let us extend the argument. Suppose there are 10 available predictors but we will only be able to afford to measure one of them in future. If we pick the one which fits best in the calibration sample, that is we pick a first-order statistic though probably it is not amongst equal variables, would Professor Lindley still use the same methods as he would if that variable had been picked independently of the calibration data?

Professor Beale has raised the topic of accuracy in least-squares calculations. The test for a singularity is much easier, not more laborious, for QR methods than for Gauss–Jordan methods. Consider the constructed data set shown in Table D1.

TABLE D1

X_1	X_2	X_3	X_4	X_5
1	1	0.47619	10.1	1
2	4	1.23809	39.9	1
3	9	2.28571	90.1	2
4	16	3.61905	159.9	2
5	25	5.23810	250.1	2
6	36	7.14286	359.9	2
7	49	9.33333	490:1	1
8	64	11.80952	639.9	1

In this example, $X_3 = X_1/3 + X_2/7$ except for rounding errors. With a column of ones inserted as a left-hand column, a QR factorization carried out in double precision gave the R-matrix shown in Table D2.

TABLE D2

	Const.	X_1	X_2	X_3	X_4	X_5
Const.	2.828	12.73	72.12	14.55	721.2	4.243
X_1		6.48	58.33	10.49	583.2	3.E–17
X_2			12.96	1.85	129.6	–1.234
X_3				8.E–6	0.011	0.309
X_4					0.276	–0.013
X_5						0.617
$\lVert X_i \rVert$	2.828	14.28	93.66	18.03	936.5	4.472

The columns contain the projections of that variable on consecutive orthogonal directions. The first direction spans the space of X_1, the second spans that part of X_2 which is orthogonal to X_1, etc. Looking down the X_3 column, we see the value 8.E–6 on the diagonal indicating that it is linearly dependent upon some combination of the previous variables, at least up to about 5 decimal places after which the variable was rounded. On the other hand, X_5 is exactly orthogonal to the space of the constant and X_1, and the value 3.E–17 is the rounding error in the computer's computation. Variable X_4 is almost 10 times X_2, and we see that after the third row the projections in the X_4 column are small. As the L_2-norm for X_4 is 936.5, about 3.5 decimal digits of cancellation error will have occurred in the calculation of the final projection of 0.276 in the X_4 column. Using a *QR* algorithm, the difference between a true singularity and high correlation is much more easily differentiated than with Gauss–Jordan methods. If we look at that part of the sum of squares and products matrix for variables X_2 and X_4, we have

$$\begin{array}{cc} 8772 & 87716.4 \\ 87716.4 & 877128.08 \end{array}$$

The elimination of variable X_2 in the Gauss–Jordan method then requires the subtraction

$$\begin{array}{r} 877128.08 \\ -877128.001477 \end{array}$$

in which the first 7 decimal digits cancel. Problems arise with *QR* algorithms because of the non-uniqueness of the Cholesky factorization when there are singularities. These problems have not been described in the literature, as far as I am aware, but there is not space here to deal with them.

I am grateful to Professor Beale for drawing my attention to the treatment of the two-variable competition case by Sprevak (1976).

The experience of Dr Kempson in adding or deleting two variables at a time is interesting. My only experience, however, with two-at-a-time algorithms has been with replacement. With a large number of available predictors, the number of combinations of 2 out of k can of course be large, but it only increases quadratically with k, not exponentially as for the exhaustive search.

Several discussants have expressed views of the kind, "If the predictors are highly correlated amongst themselves, then some of them are redundant." While this may often be true in practice, it can be completely false; the artificial example in my Table 1 illustrates a case which I have quite often encountered. The data in the paper by Fearn (1983) provide a further illustration in which all of the correlations between the predictors exceed 0.9.

Professor Gunst notes that my Table 5 does not extend to correlations of ±0.9. Below is part of the table for $\sigma = 0.3$. The strange behaviour for $\rho = +0.9$ can easily be understood by reference to Fig. 1 by imagining ellipse A stretched out so that it is very long and thin and crosses the other boundary.

	$\rho = -0.9$		$\rho = +0.9$	
β_2^*	$E(b_1 \mid sel.)$	*St. dev.*	$E(b_1 \mid sel.)$	*St. dev.*
0.0	1.03	0.27	1.08	0.34
0.5	1.10	0.23	1.03	0.37
1.0	1.23	0.19	1.07	0.40
1.5	1.41	0.15	0.96	0.74
2.0	1.62	0.13	–0.59	0.13

Dr Mallows feels that I have dismissed his work as invalid, though I cannot see this. In his paper (Mallows, 1973), he considers the case in which the subsets have been chosen independently of the data. This is the approach of many others who have derived similar formulae, such as Bendel and Afifi (1977) and Breiman and Freedman (1983), while the paper by Kohn (1983) contains asymptotic results. Formulae for the MSEP when the subsets have been chosen *a priori* would be applicable, for instance, to the fitting of polynomials or of auto-regressive models when there is a pre-determined order in which variables will be added, though once a measure such as C_p has been used as a stopping rule, some bias will be introduced. The PhD thesis of Bendel (1973) contains a considerable discussion of the various biases introduced by selection. Dr Mallows derivation makes no allowance for the terms (A) and (C) in my formula (5.2), but there is no selection bias when the subset has been chosen *a priori* so that these terms are zero in the case he considers. The reason for separately identifying selection bias is that if we can find an alternative to least-squares estimation which removes most of the selection bias then Mallows' C_p and the AIC can be used as stopping rules when the subsets of variables have not been determined independently of the data. The results of Hannan and Quinn (1979) indicate that the "$2p$" in Mallows' C_p, or more strictly the corresponding penalty for subset size in the AIC, should be $C_n p$, where C_n increases very slowly with n, say as $\log(\log n)$.

Professor Platt discusses various ways in which bootstrap methods could be applied. A paper on this subject by Professor Platt (1982) gives more details. The idea of applying the bootstrap method is appealing and seems worthy of further investigation.

The "cheap" method mentioned by Dr Green appears to be backward elimination with a lack-of-fit test. I am not clear as to whether this is used just as a pruning technique before going on to use the Beale, Kendall and Mann algorithm. The example he quotes, of fitting hyperplanes is one in which ill-conditioning can be a serious problem with some software.

Dr Preece discusses the problem of collecting "good" regression examples. In preparing this paper I made a conscious decision not to use any examples on which I had personally been involved. It is very difficult to find real problems which can be considered to be "typical" problems of a particular kind, and which do not also involve other problems which provide substantial distractions from illustrating the target technique. The examples which I have given at public lectures on subset selection in regression have often been described as untypical, for instance, they have often contained more predictors than observations. At the moment I am working with a large data set which has 9500 observations, but that has the complications of hundreds of missing values for some of the predictors and high auto-correlations. Dr Preece notes that a quadratic term is selected before the corresponding linear term for the STEAM data. In many practical cases, constraints will be applied say to stop this kind of thing happening, though it is not necessarily undesirable. I do not share his concern with the discrete nature of most of the predictors in this data set. I found the DETROIT data set far more interesting, and did do a considerable amount of experimenting with the addition of time as one of the variables, and with the incorporation of autoregressive terms, but that subset of three variables still came out a long way ahead of all that I tried.

Professor Plackett has commented that if variable selection has not been sorted out after two decades then we should move onto other problems. However, in my opinion, only a very small number of people have been looking at the statistical problems and then for only a small number of years. Most of the effort expended in this field in the past two decades has gone into computational methods, and almost all of that has neglected advances in least-squares methods which have occurred during the same period. If subset selection should be abandoned, then we should put a health warning on every statistical package which can be used for empirical model selection.

REFERENCES IN THE DISCUSSION

Bendel, R. B. (1973) *Stopping Rules in Forward Stepwise-Regression*. PhD dissertation, University of California at Los Angeles. Available from University Microfilms: Ann Arbor and London, thesis no. 74-11496.

Box, G. E. P. (1966) Use and abuse of regression. *Technometrics*, **8**, 625–629.

Breiman, L. and Freedman, D. (1983) How many variables should be entered in regression equation? *J. Amer. Statist. Ass.*, **78**, 131–136.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1983) *Classification and Regression Trees*. California: Wadsworth International Group.

Bross, I. D. J. (1982) Simplicity and credibility: a counter strategy. *Statistics and Probability Letters*, **1**, 79–83.

Clarke, M. R. B. (1982) AS 178. The Gauss-Jordan sweep operator with detection of collinearity. *Appl. Statist.*, **31**, 166–168.

Copas, J. B. (1983) Regression, prediction and shrinkage. *J. R. Statist. Soc.* B, **45**, 311–354.

Farebrother, R. W. (1980) Two empirical mean square criteria and a class of biased estimators in the standard linear model. Paper presented to Fourth World Congress of Econometric Society.

Fearn, T. (1983) A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Appl. Statist.*, **32**, 73–99.

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

Frane, J. W. (1977) A note on checking tolerance in matrix inversion and regression. *Technometrics*, **19**, 513–514.

Galpin, J. S. and Hawkins, D. G. (1982) Selecting a subset of regression variables so as to maximize the prediction accuracy at a specified point. *TWISK 275*, Tech. Report, CSIR Nat. Inst. for Math. Sci., Pretoria.

Hannan, E. J. and Quinn, B. B. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.* B, **41**, 190–195.

Hjorth, U. (1982) Model selection and forward validation. *Scand. J. Statist.*, **9**, 95–105.

——— (1983) *Model Selection Effects on Estimation and Prediction*. Unpublished report, Mathematics Department, Linkoping University, Sweden.

Hjorth, U. and Holmqvist, L. (1981) On model selection based on validation with applications to pressure and temperature prognosis. *Appl. Statist.*, **30**, 264–276.

Huang, D. Y. and Panchapakesan, S. (1982) On eliminating inferior regression models. *Commun. in Statist.*, **A11**, 751–759.

Lindley, D. V. (1968) The choice of variables in multiple regression (with Discussion). *J. R. Statist. Soc.* B, **30**, 31–66.

Mabbett, A., Stone, M. and Washbrook, J. (1980) Cross-validatory selection of binary variables in differential diagnosis. *Appl. Statist.*, **29**, 198–204.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

O'Brien, C. M. and Stone, M. (1984) A new procedure to analyse radioactive emission count data. *Nuclear Instruments and Methods in Physics Research*, **218**, 130–136.

Platt, C. A. (1982) Bootstrap stepwise regression. *Proc. Bus. and Econ. Sect., Amer. Statist. Assoc.*, 586–589.

Sprevak, D. (1976) Statistical properties of estimates of linear models. *Technometrics*, **18**, 283–289.