The Annals of Mathematical Statistics
1971, Vol. 42, No. 4, 1273–1284

MODEL BUILDING FOR PREDICTION IN REGRESSION BASED UPON REPEATED SIGNIFICANCE TESTS[1]

By W. J. Kennedy and T. A. Bancroft
Iowa State University

1. Introduction.

1.1. Regression analysis considered herein will deal with the fitting of linear models for the purpose of prediction. The method of least squares provides a well-defined mathematical procedure for obtaining a unique prediction equation whenever it can reasonably be assumed that the data arose from a situation which can adequately be represented by a linear model having one dependent variable and a definite number of independent variables. The usual additional assumption of independently normally distributed errors having zero mean and a constant variance allows application of additional statistical theory to test selected hypotheses of interest and to set confidence intervals.

In application of the theory of regression analysis to experimental data, uncertainty often arises as to the exact number of independent variates to include in the final model. Many situations are such that among the total set of independent variates only a small subset is of real value when attempting to predict the behavior of the dependent variable.

Several different procedures have been recommended for use in determining a suitable subset of independent variables for use in predicting the dependent variable of interest (see Abt [1], Draper and Smith [9], Efroymson [10], Gorman and Toman [12], Hocking and Leslie [13]). These procedures involve the use of repeated tests of significance and rely upon inferences based upon the outcome of such tests. The decision rules used in these procedures were, for the most part, selected for their intuitive appeal and little consideration has been given to the consequences, with respect to the fitted model, of the effect on subsequent inferences of such repeated testing.

The problem of model building in regression is one of the general class of problems called problems of incompletely specified models involving the use of repeated tests of significance. This classification serves to clarify the nature of this regression problem and to emphasize the need for more relevant theoretical development in this problem area. The development of model building techniques in general has been carried out under the assumption that no *a priori* information is available to the experimenter concerning which variables should remain in the final model.

1.2. *Objectives of the present study*. The present study will concern itself with two different model building procedures, called "Forward Selection" and "Sequential Deletion". We will consider these for use in model building when the experimenter believes that the usual error assumptions are appropriate in his full regression

Received March 19, 1970; revised January 4, 1971.

[1] This research was partially supported by the National Science Foundation Grant GP9046.

model, however, he feels that many of the independent variates will be of no real value in a predictor of the predictand y.

Each of the two procedures assumes that the experimenter has sufficient knowledge in the area of application to allow him to select the "basic set" of r independent variables (which could be the null set) and to designate an "order of importance" for the remaining $k-r$ variables with x_{r+1} being most important and x_k least important. In many cases such knowledge is gained from theoretical considerations or from a substantial amount of experience in the applied area. In other cases, such as polynomial regressions, a natural order is given.

The model building procedures are described by the following:

(Sequential Deletion). The experimenter has n measurements available on the $k+1$ variables $(y, x_1, \cdots, x_k)$. He wants to predict y on the basis of the values the x_i will assume. He first tests the hypothesis that the coefficient of x_k is zero (i.e., x_k is not needed in the equation). If he accepts this hypothesis he deletes x_k and tests that the coefficient of x_{k-1} is zero. If he accepts the second hypothesis he deletes x_{k-1} from the prediction equation and tests the coefficient of x_{k-2}, etc. He continues deleting variables in this manner until he rejects a hypothesis that a coefficient is zero, or until he reaches the coefficient of $x_r (r < k)$, then he retains in his prediction equation the variable corresponding to that coefficient and all other variables whose coefficients he has not yet tested.

(Forward Selection). The experimenter has n measurements available on the $k+1$ variable $(y, x_1, x_2, \cdots, x_k)$. He also assumes that the first $r (r < k)$ of the k independent variables are necessary for prediction of y. He then tests the hypothesis that the coefficient of x_{r+1} is zero. If he rejects this hypothesis he adds x_{r+1} to the list of necessary variables and tests that the coefficient of x_{r+2} is zero. If he rejects this second hypothesis he adds x_{r+2} to the list of necessary variables and tests that the coefficient of x_{r+3} is zero, etc. He continues adding variables to his prediction equation in this manner until he arrives at a variable whose coefficient does not differ significantly from zero, at which point he does not add that variable to the equation, nor does he add the variables whose coefficients he has not yet tested.

The objectives of the present study are threefold. First, to provide a means for examining, with respect to bias and mean square error of predictand, the consequences of using the two model building procedures. The second objective is to give a summary of the results obtained in a numerical study of the efficiency of the two procedures relative to one another and to the procedure wherein all independent variates are retained. Finally, to recommend, based upon the results of the numerical study, significance levels for use in model building in various circumstances.

1.3. *Related papers.* Previous, related papers which deal with the problems of incompletely specified models involving the use of single and repeated tests or preliminary tests of significance include investigations by Bancroft [5], Bechhofer [6], Bennett [7], Bozovich, Bancroft, and Hartley [8], Huntsberger [14], Kitagawa [16], Mead [19], and Paull [20]. Publications which deal more specifically with the theory of model building in regression analysis using sequential testing include Anderson [2], Anscombe [3], and Larson and Bancroft [17], [18].

2. The sequential deletion procedure. This procedure, as previously described, can be summarized mathematically as follows. (The estimator of the true value of y for any case is denoted by y^*; the superscript on y denotes the number of independent variables included in the fitted prediction equation.)

Event	y^*	Situation
A_i	y_{k-i}	Reject H_i: $\beta_{k-i}=0$; Accept H_{i-1}:
		$\beta_{k-i+1}=0;\cdots$ Accept $H_1:\beta_{k-1}=0$;
		Accept $H_0:\beta_k=0$. $(i=0,1,2,\cdots,k-r-1)$.
A_{k-r}	y_r	Accept $H_{k-r-1}:\beta_{r+1}=0$; Accept H_{k-r-2};
		$B_{r+2}=0;\cdots$ Accept $H_1:\beta_{k-1}=0$;
		Accept $H_0:\beta_k=0$.

This mathematical form also serves to emphasize the fact that the data themselves are used as a basis for determining which of the $k-r+1$ linear models to use in predicting values of y. We will assume throughout that all repeated tests are made at the same level α.

2.1. *The bias in y^*.* Assume that the true model generating our data is $Y = X\beta+e$ where Y is the $n\times 1$ vector of observed y values, X is the $n\times(k+1)$ matrix of x values, and e is the $n\times 1$ vector of error components. We assume that $E(ee') = \sigma^2 I$, that the independent variates are transformed so that the columns with the exception of the first column have mean zero, and that $X'X = I$. (In Section 5 we show that the bias and mean square error of y^* is not affected by this transformation.)

The test criterion for the hypothesis $\beta_i = 0$ is b_i^2/v, $(i = r+1, \cdots, k)$ where b_i is the least squares estimate of β_i and v is the residual mean square obtained by fitting the full model. The hypothesis is rejected if $b_i^2/v \geqq \delta$ (the $100(1-\alpha)$ percent point of $F_{1,n-k-1}$) and accepted otherwise. Then the expected value of y^* is

$$(1)\quad E(y^*) = E(y_k \mid A_0)P(A_0)+E(y_{k-1} \mid A_1)P(A_1)+\cdots+E(y_r \mid A_{k-r})P(A_{k-r}),$$

where, due to the fact that the b_i are independent, we have for $i = r+1, r+2, \cdots, k$ that

$$E(y_i \mid A_{k-i})P(A_{k-i}) = [\beta_0+\beta_1 x_1+\cdots+\beta_{i-1}x_{i-1}+x_i E(b_i \mid A_{k-i})]P(A_{k-i})$$

and

$$E(y_r \mid A_{k-r}) = (\beta_0+\textstyle\sum_{j=1}^{r}\beta_j x_j)P(A_{k-r}).$$

We introduce the notation $\lambda_i = \beta_i^2/2\sigma^2$, $m = n-k-1$, and use $F_s(z \mid \lambda)$ to denote the cumulative distribution function of the noncentral chi square having s degrees of freedom and noncentrality λ. For any $0 \leqq i < k-r$ the probability $P(A_i)$ is expressible in the form

$$(2)\qquad C_{k-i} = P(A_i) = \int_0^\infty [\textstyle\prod_{j=k-i+1}^{k} F_1(\gamma y \mid \lambda_j)][1-F_1(\gamma y \mid \lambda_{k-i})]g(y)\,dy$$

where $\gamma = \delta/m$, $g(y)$ is the p.d.f. of ${\chi_m}^2$, and we define $\prod_{j=k+1}^{k} F_1(x \mid \lambda_j) \equiv 1$. If we let $r(b_{k-i}, b_{k-i+1}, \cdots, b_k, v)$ denote the joint density of the independent random variables $b_{k-i}, b_{k-i+1}, \cdots, b_k, v$ then we have

$$P(A_i) = \int_B \cdots \int r(b_{k-i}, \cdots, b_k, v)\, dv \prod_{j=k-i}^{k} db_j \tag{3}$$

where B is the region defined by $\{y: {b_k}^2 < \delta v; \cdots; b_{k-i+1}^2 < \delta v; b_{k-i}^2 \geqq \delta v\}$. Differentiating C_{k-i} with respect to β_{k-i} gives

$$\frac{\partial C_{k-i}}{\partial \beta_{k-i}} = \frac{\beta_{k-i}}{\sigma^2}[H(A_i) - P(A_i)]$$

where

$$H(A_i) = \int_0^\infty \left[\prod_{j=k-i+1}^{k} F_1(\gamma y \mid \lambda_j)\right][1 - F_3(\gamma y \mid \lambda_{k-i})]g(y)\, dy.$$

Differentiating (3) with respect to β_{k-i} gives

$$\frac{\partial P(A_i)}{\partial \beta_{k-i}} = \frac{1}{\sigma^2} E(b_{k-i} \mid A_i) P(A_i) - \frac{\beta_{k-i}}{\sigma^2} P(A_i).$$

Equating the two derivatives gives

$$E(b_{k-i} \mid A_i) P(A_i) = \beta_{k-i} H(A_i) \qquad (0 \leqq i < k - r). \tag{4}$$

Using Equation (4) we substitute into Equation (1) to obtain $E(y^*)$. Using the convention $\sum_{s=0}^{-1} P(A_s) \equiv 0$, the bias in y^* is then expressible as

$$\text{bias}(y^*) = \sum_{t=r+1}^{k} \beta_t x_t \left[\sum_{s=0}^{k-t-1} P(A_s) + H(A_{k-t}) - 1\right]. \tag{5}$$

2.2. *The mean square error of y^*.* Having derived an expression for $E(y^*)$ we need $E[(y^*)^2]$ in order to obtain the mean square error of y^*. Using previously defined relationships between estimators and events it is easily seen that

$$E(y^*)^2 = E({y_k}^2 \mid A_0)P(A_0) + E(y_{k-1}^2 \mid A_1)P(A_1) + \cdots + E({y_r}^2 \mid A_{k-r})P(A_{k-r}). \tag{6}$$

Expanding these terms we have

$$\begin{aligned} E({y_i}^2 \mid A_{k-i})P(A_{k-i}) = {} & [(\beta_0 + \textstyle\sum_{j=1}^{i-1} \beta_j x_j)^2 + \sigma^2(1/n + \sum_{j=1}^{i-1} {x_j}^2) \\ & + 2(\beta_0 + \textstyle\sum_{j=1}^{i-1} \beta_j x_j) x_i E(b_i \mid A_{k-i}) + {x_i}^2 E({b_i}^2 \mid A_{k-i})]P(A_{k-i}), \end{aligned}$$

for $r < i \leqq k$ and

$$E({y_r}^2 \mid A_{k-r})P(A_{k-r}) = [(\beta_0 + \textstyle\sum_{j=1}^{r} \beta_j x_j)^2 + \sigma^2(1/n + \sum_{j=1}^{r} {x_j}^2)]P(A_{k-r}). \tag{7}$$

The expectation of the ${y_i}^2$ depends on that of ${b_i}^2$. To derive a usable expression for $E({b_i}^2 \mid A_{k-i})P(A_{k-i})$ we proceed as follows. First differentiate $P(A_i)$ in (3) twice with respect to β_{k-i}^2. The result is easily seen to be

$$\frac{\partial^2 P(A_i)}{\partial \beta_{k-i}^2} = \frac{1}{\sigma^4}\{E(b_{k-i}^2 \mid A_i) - 2\beta_{k-i} E(b_{k-i} \mid A_i) + \beta_{k-i}^2 - \sigma^2\} P(A_i).$$

Similarly, differentiating C_{k-i} we obtain

$$\frac{\partial^2 C_{k-i}}{\partial \beta_{k-i}^2} = \frac{\beta_{k-i}^2}{\sigma^4}[P(A_i) - 2H(A_i) + T(A_i)] + \frac{1}{\sigma^2}[H(A_i) - P(A_i)]$$

where

$$T(A_i) = \int_0^\infty [\prod_{j=k-i+1}^{k} F_1(\gamma y \mid \lambda_j)][1 - F_5(\gamma y \mid \lambda_{k-i})]g(y)\,dy.$$

Again using the equality of partial derivatives we obtain

$$E(b_{k-i}^2 \mid A_i)P(A_i) = \beta_{k-i}^2 T(A_i) + \sigma^2 H(A_i).$$

Substitution into (6) gives, upon simplification,

$$\begin{aligned} E(y^*)^2 = {} & \sum_{j=r+1}^{k} \{(\beta_0 + \sum_{i=1}^{j-1} \beta_i x_i)^2 + \sigma^2(1/n + \sum_{i=1}^{j-1} x_i^2)\} P(A_{k-j}) \\ & + 2\sum_{j=r+1}^{k} (\beta_0 + \sum_{i=1}^{j-1} \beta_i x_i) x_j \beta_j H(A_{k-j}) + \sum_{j=r+1}^{k} x_j^2 [\beta_j^2 T(A_{k-j}) \\ & + \sigma^2 H(A_{k-j})] + [(\beta_0 + \sum_{i=1}^{r} \beta_i x_i)^2 + \sigma^2(1/n + \sum_{i=1}^{r} x_i^2)] P(A_{k-r}). \end{aligned} \tag{8}$$

We now obtain the mean square error of y^* as

$$\begin{aligned} \text{mse}(y^*) &= E(y^*)^2 - [E(y^*)]^2 + [\text{bias}(y^*)]^2 \\ &= \sum_{j=r+1}^{k} [(\sum_{i=r+1}^{j-1} \beta_i x_i)^2 P(A_{k-j}) - 2(\sum_{i=r+1}^{k} \beta_i x_i)\beta_j x_j \sum_{s=0}^{k-j-1} P(A_s) \\ &\quad + \sigma^2(1/n + \sum_{i=1}^{j-1} x_i^2) P(A_{k-j}) + (\sigma^2 x_j^2 - 2\beta_j x_j \sum_{i=1}^{k} \beta_i x_i) H(A_{k-j}) \end{aligned} \tag{9}$$

$$+ \beta_j^2 x_j^2 T(A_{k-j})] + \sigma^2(1/n + \sum_{i=1}^{r} x_i^2) P(A_{k-r}) + (\sum_{i=r+1}^{k} \beta_i x_i)^2, \tag{10}$$

where $\sum_{i=k}^{k-1} \beta_i x_i \equiv 0$.

There are partial checks which can easily be shown to hold for this expression. If $\delta = 0$ corresponding to always rejecting H_0, the mean square error is known to be $\sigma^2(1/n + \sum_{i=1}^{k} x_i^2)$. As $\delta \to \infty$ corresponding to always using only the first r variates to predict y, the mean square error is $\sigma^2(1/n + \sum_{i=1}^{r} x_i^2) + (\sum_{i=r+1}^{k} \beta_i x_i)^2$. Finally, this mean square error agrees with that obtained by Larson and Bancroft [17] for the case of $r = k-1$.

3. The forward selection procedure. This procedure, as previously described, can be summarized mathematically as follows:

Event	y^*	Situation
A_i	y_{r+i}	Accept $H_i: \beta_{r+i+1} = 0$, Reject $H_{i-1}: \beta_{r+i} = 0, \cdots$
		Reject $H_0: \beta_{r+1} = 0, i = 0, 1, \cdots, k-r-1$,
A_{k-r}	y_k	Reject $H_{k-r-1}: \beta_k = 0$; Reject $H_{k-r-2}: \beta_{k-1} = 0; \cdots$
		Reject $H_0: \beta_{r+1} = 0$.

As was the case in sequential deletion, the forward selection procedure includes a sequential testing of hypotheses and the particular model arrived at through use of this procedure is determined by the results of these tests.

3.1. *The bias in y^*.* We make the same assumptions about the linearity of model, the orthogonality of x vectors, and the distribution of errors, as were made in Section 2.1. Again the assumption of orthogonality of x vectors is not restrictive because an appropriate transformation to non-orthogonal data does not alter the bias and mean square error of predicted y.

The criterion used in testing the hypothesis $\beta_i = 0$ is b_i^2/v $(i = r+1, r+2, \cdots, k)$, where v is the residual mean square from the full fit. The expected value of y^* is expressible as

$$E(y^*) = (\beta_0 + \textstyle\sum_{i=1}^{r} \beta_i x_i) + \sum_{i=1}^{k-r} \sum_{j=r+1}^{r+i} x_j E(b_j \mid A_i) P(A_i). \tag{11}$$

In order to obtain a more informative expression for $E(b_j \mid A_i)P(A_i)$ we again express $P(A_i)$ in integral form as

$$C_i = P(A_i) = \int_0^\infty \prod_{j=1}^{i} [1 - F_1(\gamma y \mid \lambda_{r+j})] F_1(\gamma y \mid \lambda_{r+i+1}) g(y)\, dy, \tag{12}$$

where

$$\prod_{j=1}^{0} [1 - F_1(\gamma y \mid \lambda_{r+j})] \equiv 1.$$

Alternatively we have

$$P(A_i) = \int \cdots \int_B r(b_{r+1}, \cdots, b_{r+i+1}, v) \prod_{j=r+1}^{r+i+1} db_j\, dv \tag{13}$$

where B is the region defined by $\{y: b_{r+1}^2 \geqq \delta v;\ b_{r+2}^2 \geqq \delta v;\ \cdots;\ b_{r+i} \geqq \delta v;\ b_{r+i+1} < \delta v\}$. For any integer t in the range $1 \leqq t \leqq i$ we differentiate $P(A_i)$ in (13) with respect to β_{r+t} and obtain

$$\frac{\partial P(A_i)}{\partial \beta_{r+t}} = \frac{1}{\sigma^2} E(b_{r+t} - \beta_{r+t} \mid A_i) P(A_i).$$

Using (12) we again differentiate with respect to β_{r+t} and obtain

$$\frac{\partial C_i}{\partial \beta_{r+t}} = \frac{\beta_{r+t}}{\sigma^2} [H_{r+t}(A_i) - P(A_i)]$$

where

$$H_{r+t}(A_i) = \int_0^\infty \prod_{j=1;j\neq t}^{i} [1 - F_1(\gamma y \mid \lambda_{r+j})][1 - F_3(\gamma y \mid \lambda_{r+t})] F_1(\gamma y \mid \lambda_{r+i+1}) g(y)\, dy.$$

Using equality of these partial derivatives gives $E(b_{r+t} \mid A_i)P(A_i) = \beta_{r+t} H_{r+t}(A_i)$, and substituting into (10) we have

$$E(y^*) = (\beta_0 + \textstyle\sum_{i=1}^{r} \beta_i' x_i) + \sum_{i=1}^{k-r} \sum_{j=r+1}^{r+i} \beta_j x_j H_j(A_i). \tag{14}$$

Thus, the bias is expressible in the form

$$\text{bias}(y^*) = \sum_{i=1}^{k-r} \sum_{j=r+1}^{r+i} \beta_j x_j H_j(A_i) - \sum_{i=r+1}^{k} \beta_i x_i. \tag{15}$$

3.2. *The mean square error.* We will use the general approach employed in Section 2.2 to obtain mse (y^*). As before we have

$$(16)\quad E(y^*)^2 = E(y_r{}^2 \mid A_0)P(A_0)+E(y_{r+1}^2 \mid A_1)P(A_1)+\cdots+E(y_k{}^2 \mid A_{k-r})P(A_{k-r}).$$

Considering individual terms in this sum we see that

$$\begin{aligned}(17)\quad E(y_i{}^2 \mid A_{i-r})P(A_{i-r}) &= [(\beta_0+\textstyle\sum_{m=1}^{r}\beta_m x_m)^2+\sigma^2(1/n+\sum_{m=1}^{r}x_m{}^2)\\ &\quad +2(\beta_0+\textstyle\sum_{m=1}^{r}\beta_m x_m)\sum_{j=r+1}^{i}x_j E(b_j \mid A_{i-r})\\ &\quad +2\textstyle\sum_{j=r+2}^{i}\sum_{m=r+1;j<m}^{i-1}x_j x_m E(b_j b_m \mid A_{i-r})\\ &\quad +\textstyle\sum_{j=r+1}^{i}x_j{}^2 E(b_j{}^2 \mid A_{i-r})]P(A_{i-r}),\ i=r+1,r+2,\cdots,k.\end{aligned}$$

This equation points to the need for expressing $E(b_i b_j \mid A_t)P(A_t)$ in some more usable form. First consider the case where $i = j$. Taking second partial derivatives we have

$$-\frac{\partial^2 P(A_t)}{\partial\beta_i{}^2}=\frac{1}{\sigma^2}P(A_t)-\frac{1}{\sigma^4}E[(b_i-\beta_i)^2 \mid A_t)]P(A_t).$$

$$\frac{\partial^2 C_t}{\partial\beta_i{}^2}=\frac{\beta_i{}^2}{\sigma^4}[P(A_t)-2H_i(A_t)+T_i(A_t)]+\frac{1}{\sigma^2}[H_i(A_t)-P(A_t)]$$

where

$$T_i(A_t)=\int_0^\infty \textstyle\prod_{s=r+1;s\neq i}^{r+t}[1-F_1(\gamma y \mid \lambda_s)][1-F_5(\gamma y \mid \lambda_i)]F_1(\gamma y \mid \lambda_{r+t+1})g(y)\,dy$$

$i \leqq r+t$, $F_1(\gamma y \mid \lambda_{k+1}) \equiv 1$. Equating the two derivatives and using the expression for $E(b_i \mid A_t)P(A_t)$ given in Section 3.1 we obtain

$$(18)\qquad E(b_i{}^2 \mid A_t)P(A_t) = \beta_i{}^2 T_i(A_t)+\sigma^2 H_i(A_t).$$

For the case of $i \neq j$ we use the following derivatives.

$$\frac{\partial^2 P(A_t)}{\partial\beta_i\,\partial\beta_j}=\frac{1}{\sigma^4}E[(b_i-\beta_i)(b_j-\beta_j) \mid A_t]P(A_t),$$

$$\frac{\partial^2 C_t}{\partial\beta_i\,\partial\beta_j}=\frac{\beta_i\beta_j}{\sigma^4}[S_{ij}(A_t)-H_i(A_t)-H_j(A_t)+P(A_t)],$$

where

$$S_{ij}(A_t)=\int_0^\infty \textstyle\prod_{s=r+1;s\neq i,j}^{r+t}[1-F_1(\gamma y \mid \lambda_s)][1-F_3(\gamma y \mid \lambda_i)][1-F_3(\gamma y \mid \lambda_j)]\cdot F_1(\gamma y \mid \lambda_{r+t+1})g(y)\,dy.$$

Using the equality of the two derivatives we have

$$(19)\qquad E(b_i b_j \mid A_t)P(A_t) = \beta_i\beta_j S_{ij}(A_t).$$

The results given by (14), (15), (18), and (19) are now used to obtain

$$\begin{aligned}(20)\quad \text{mse}\,(y^*) &= E(y^*)^2-[E(y^*)]^2+[\text{bias}(y^*)]^2\\ &= \sigma^2(1/n+\textstyle\sum_{i=1}^{r}x_i{}^2)+\sum_{j=1}^{k-r}\{\sum_{m=r+1}^{r+j}x_m{}^2[\beta_m{}^2 T_m(A_j)+\sigma^2 H_m(A_j)]\\ &\quad +2\textstyle\sum_{i=r+1}^{j}\sum_{t=r+2;i<t}^{j}\beta_i\beta_t x_i x_t S_{it}(A_j)\}\\ &\quad -2(\textstyle\sum_{i=r+1}^{k}\beta_i x_i)\sum_{i=1}^{k-r}\sum_{j=r+1}^{r+i}\beta_j x_j H_j(A_i)+(\sum_{i=r+1}^{k}\beta_i x_i)^2.\end{aligned}$$

The partial checks used for Sequential Deletion can also be shown to hold in the right member of (20).

4. Relative efficiency. The final stated objective of this study is to investigate the efficiency of the two model building procedures. We will first define relative efficiency for each of the procedure pairs to be studied.

4.1. *Definitions.* We define the relative efficiency R_0 of the forward selection to the sequential deletion procedure to be, for a common set of n observations,

$$R_0 = \frac{M_1}{M_2} = \frac{n/\sum_{\text{obs}}(\text{mse}_i(y^*) \text{ for forward selection})}{n/\sum_{\text{obs}}(\text{mse}_i(y^*) \text{ for sequential deletion})} \tag{21}$$

where $(1/n)\sum_{\text{obs}} \text{mse}_i(y^*)$ denotes the average of the mean square errors averaging over the n observations. Here M_1 and M_2 are respectively the simplified form of the numerator and denominator of R_0. These average values used in defining R_0 are estimates of the respective population mean square errors of predicted y over the space of x's. Using the previously derived expressions (10) and (20) along with the properties of the X matrix we obtain R_0 in the form[2] $R_0 = M_1/M_2$ where

$$\begin{aligned} M_1 &= \textstyle\sum_{t=r+1}^{k} [tP(A_{k-t}) - 2\lambda_t \sum_{s=0}^{k-t-1} P(A_s) + (1-4\gamma_t)H(A_{k-t}) + 2\lambda_t T(A_{k-t})] \\ &\quad + (r+1)P(A_{k-r}) + 2\textstyle\sum_{j=r+1}^{k} \lambda_j \qquad (22) \\ M_2 &= r+1+\textstyle\sum_{j=1}^{k-r} \sum_{m=r+1}^{r+j} [2\lambda_m T_m(A_j) + (1-4\lambda_m)H_m(A_j)] + 2\sum_{j=r+1}^{k} \lambda_j. \end{aligned}$$

The relative efficiency of forward selection to the procedure wherein all independent variables are always retained in the model (called the "always keep" procedure) is defined to be

$$R_1 = \frac{n/\sum_{\text{obs}}(\text{mse}_i(y^*) \text{ for forward selection})}{n/\sum_{\text{obs}}(\text{mse}_i(y^*) \text{ for always keep})}.$$

The ratio R_1 is easily seen to be $(k+1)/M_2$. Using the same form of definition for the relative efficiency of Sequential Deletion to the "always keep" procedure the definition gives rise to $R_2 = (k+1)/M_1$.

4.2. *Results of numerical study.* An extensive numerical study was made in order to compare the two model building procedures and to determine at which level α might best be used for tests made in model building[3]. The range of parameters used was

[n: 20, 60, others in special cases], [k: 5, 10, 20], [λ_i; 0(1)5],
[$k-r$: 1(1)5].

For each combination of the parameters the value of R_0, R_1 and R_2 was obtained at each of the levels $\alpha = 0.50, 0.25, 0.10, 0.05$. Other combinations of these parameters were considered in regions where a "finer grid" appeared to be desirable.

---

[2] Note that σ^2 is now included in the λ_i.

[3] The results obtained in computation are on file in the Statistical Laboratory at Iowa State University, Ames, Iowa.

The ratio R_0 was used for comparison of the two model building procedures. The values of R_0 were generally in the range $0.50 \leqq R_0 \leqq 1.20$ with some exceptions. The majority of the R_0 values obtained were less than 0.95 in value and the mean R_0 was 0.86. Figure 1 is a representative example of the results obtained. The numerical values obtained indicated clearly that the forward selection procedure is relatively less efficient than the sequential deletion procedure for all cases except those where the parameters λ_i are all very small.

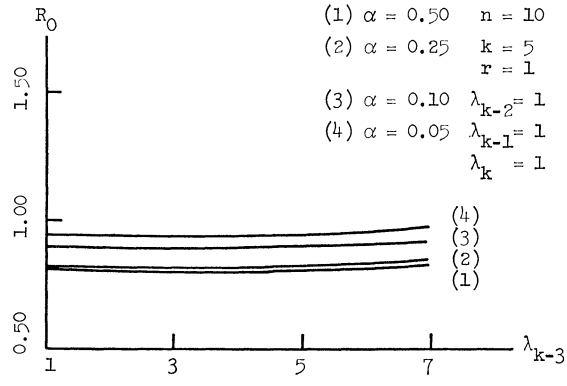

FIG. 1. Representative graph of R_0 as a function of λ_{k-3}.

The ratios R_1 and R_2 were used to study relative efficiency in relation to test levels α. Results obtained proved that it is not possible to find a level α which can assure that the relative efficiency R_1 or R_2 is maintained above some nominal level (e.g., 0.80) for all possible combinations of the other parameter values. This result was expected and the study was made primarily to see which levels α are most appropriate according to the relative efficiency criterion. The majority of values obtained in the study were in the ranges $0.70 \leqq R_1 \leqq 1.15$, and $0.70 \leqq R_2 \leqq 1.05$. The approximate median values for the R_1 and R_2 were 0.83 and 0.92, respectively. No single α level was found to be universally superior. The conclusion reached, based upon the numerical results obtained, is that a good choice of α is one in the range $0.10 \leqq \alpha \leqq 0.25$. Specifically for R_1 a level α near 0.25 seems best and for R_2 an α near 0.10 seems best. Figure 2 is representative[4] of the numerical results obtained.

In addition to the numerical study described above, the expressions derived by Larson and Bancroft [18] for mean square error of predicted y when σ^2 is known were used to define the analogues R_0', R_1', R_2' of R_0, R_1, R_2 and these ratios were used in a second numerical study. The range of values used in this study was

$$[k - r: 2(1)5], [k: 5(5)25, 50(25)100],$$
$$[\beta_i/\sigma: 0(1)5 \text{ for } k - r = 2, 3, \text{ and } 0(1)3 \text{ for } k - r = 4, 5].$$

[4] The word "representative" is used to indicate that the majority of the graphs had this general appearance.

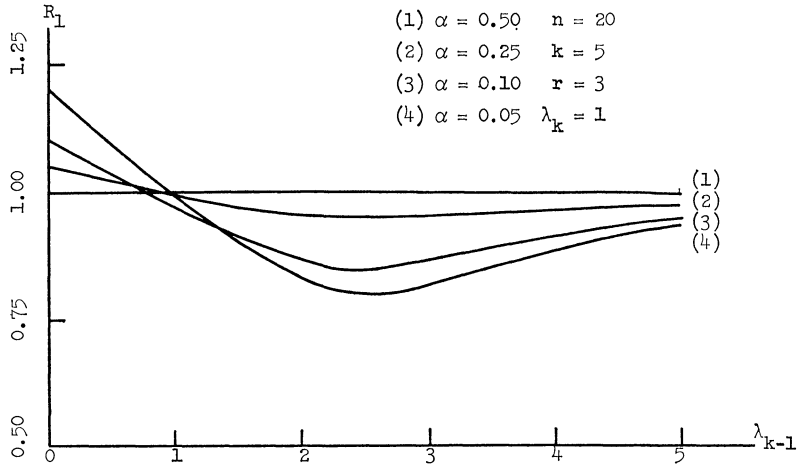

FIG. 2. Representative graph of R_1 as a function of λ_{k-1}.

The results were much the same as those previously summarized for the case of unknown σ^2, and the same conclusions were reached. This constitutes an extension of the results obtained by Larson and Bancroft for the case of σ^2 known.

Based upon the numerical results obtained the sequential deletion procedure is recommended for use. The level $\alpha = 0.25$ for tests made in application of this procedure appears to be very appropriate. The forward selection procedure is generally less efficient and cannot be recommended over sequential deletion.

5. The nonorthogonal case. The preceding derivations of bias and mean square error were made using a transformed X matrix. The transformation was made to achieve orthogonal x's. In this section we show that the bias and mean square error of predicted y is the same for original and transformed x's. Larson and Bancroft [17] have derived similar results for the case of known error variance σ^2.

5.1. *Equality of biases.* The proof will be given for the Sequential Deletion procedure. The corresponding result for Forward Selection is derived in the same way. Let X_{k-i} $(i = 0, 1, \cdots, k-r)$ be the $n \times (k-i)$ matrix composed of the first $k-i$ columns of the original X matrix. Let b_{k-i} $(i = 0, 1, \cdots, k-r)$ be the $(k-i) \times 1$ vector of regression coefficients on X_{k-i}. Let ϕ_k be an arbitrary $1 \times k$ vector and ϕ_{k-i} $(i = 0, \cdots, k-r)$ be the $1 \times (k-i)$ vector containing the first $k-i$ components of ϕ_k. The bias in the estimator of $\phi_k\beta$ is denoted by $g(X_k, X_{k-1}, \cdots, X_r; \phi_k; \sigma^2)$. Using the basic assumption about the distribution of Y we have that $g(X_k, X_{k-1}, \cdots, X_r; \beta; \phi_k; \sigma^2)$ is given as follows.

Define Δ_j $(j = r+1, \cdots, k)$ to be $\Delta_j = (R(X_j; Y) - R(X_{j-1}; Y))/v$. Also, define the Ω_i $(i = 1, 2, \cdots, k-r+1)$ to be

$$\Omega_1 = \{Y : \Delta_k \geqq \delta\},$$

$$\Omega_j = \{Y : \Delta_{k-j+1} \geqq \delta, \Delta_{k-j+2} < \delta, \cdots, \Delta_k < \delta\}, \quad (j = 2, \cdots, k-r),$$

$$\Omega_{k-r+1} = \{Y : \Delta_{r+1} < \delta, \Delta_{r+2} < \delta, \cdots, \Delta_k < \delta\}.$$

Then we have

$$(23)\quad g(X_k, X_{k-1}, \cdots, X_r; \beta; \phi_k; \sigma^2) = \sum_{j=1}^{k-r+1} \int_{\Omega_j} \phi_{k-j+1} b_{k-j+1} \cdot N(Y; X\beta; \sigma^2 I) \prod_{i=1}^{n} dy_i - \phi_k \beta$$

where $R(X_{k-i};\ Y)$ is the regression sum of squares obtained by regressing Y on X_{k-i}, and $N(Y;\ X\beta;\ \sigma^2 I)$ is the probability density function of the multivariate normal distribution having mean vector $X\beta$ and variance-covariance matrix $\sigma^2 I$.

There exists a nonsingular upper triangular matrix A_k such that $(XA_k)'(XA_k) = I$. (Several different orthogonalization procedures will give A; the well-known Gram–Schmidt procedure is one such procedure.) Let A_{k-i} be the $(k-i)\times(k-i)$ principal submatrix of A_k $(i = 0, 1, \cdots, k-r)$. Let $Z_k = XA_k$ denote the transformed X matrix and $Z_{k-i} = X_{k-i}A_{k-i}$ $(i = 0, 1, \cdots, k-r)$ denote the first $k-i$ columns of the Z matrix. Using this notation the following relationships are easily seen to hold.

(1) $R(X_{k-i};\ Y) = R(X_{k-i}A_{k-i};\ Y)$ for every $i = 0, 1, \cdots, k-r$.

(2) $b_{k-i} = A_{k-i}d_{k-i}$ for each $i = 0, 1, \cdots, k-r$ where d_{k-i} is the $(k-i)\times 1$ vector of regression coefficients on $X_{k-i}A_{k-i}$.

(3) $\phi_{k-i}b_{k-i} = \phi_{k-i}A_{k-i}d_{k-i}$ for each $i = 0, 1, \cdots, k-r$.

Using (23) and these three relationships it follows that

$$(24)\quad g(X_k, X_{k-1}, \cdots, X_r; \beta; \phi_k; \sigma^2) = g(X_kA_k, X_{k-1}A_{k-1}, \cdots, X_rA_r; A_k^{-1}\beta; \phi_kA_k; \sigma^2).$$

Thus, (24) shows that the bias in y^* is the same for original and transformed x variates whenever the Sequential Deletion procedure is used. The same is true for Forward Selection and the proof follows readily using an argument similar to the one given above.

5.2. *Equality of mean square errors.* An expression for the variance of predicted y is obtainable by rewriting (24) with slighly modified integrands. Using the relationships (1), (2), and (3) it follows immediately that the variance of y^*, and, hence, the mean square error of y^*, is not affected by the use of transformed x variates.

6. Acknowledgment. The authors wish to thank both the referee and associate editor whose helpful comments led to improvements in the original manuscript. Thanks also go to Professor Chien Pai Han for his participation in many enlightening discussions on the subject of this paper.

REFERENCES

[1] ABT, K. (1967). Significant independent variables in linear models. *Metrika* **12** 2–15.

[2] ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **22** 255–266.

[3] ANSCOMBE, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *J. Roy. Statist. Soc. Ser. B* **29** 1–59.

[4] BANCROFT, T. A. (1950). Bias due to the omission of independent variables in ordinary multiple regression analysis (abstract). *Ann. Math. Statist.* **21** 142.

[5] BANCROFT, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Statist.* **15** 190–204.

[6] BECHHOFER, R. E. (1951). The effect of preliminary tests of significance on the size and power of certain tests of univariate linear hypothesis. Ph.D. thesis, Columbia Univ.

[7] BENNETT, B. M. (1955). On the use of preliminary tests in certain statistical procedures. *Ann. Inst. Statist. Math.* **8** 45–52.

[8] BOZIVICH, H., BANCROFT, T. A. and HARTLEY, H. O. (1956). Power of analysis of variance test procedures for certain incompletely specified models, I. *Ann. Math. Statist.* **27** 1017–1043.

[9] DRAPER, N. and SMITH, H. (1966). Applied regression analysis. Wiley, New York.

[10] EFROYMSON, M. A. (1964). Multiple regression analysis. *Mathematical Methods for Digital Computers* (A. Ralston and H. Wilf, eds.). Wiley, New York.

[11] GARSIDE, M. J. (1965). The best subset in multiple regression analysis. *Appl. Statist. Ser. C* **14** 196–200.

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

[13] HOCKING, R. R. and LESLIE, R. N. (1967). Best subset in regression analysis. *Technometrics* **9** 531–540.

[14] HUNTSBERGER, D. V. (1955). A generalization of a preliminary testing procedure for pooling data. *Ann. Math. Statist.* **26** 734–643.

[15] KENNEDY, W. J. (1969). Model building for prediction in regression analysis based on repeated significance tests. Ph.D. thesis, Iowa State Univ. of Science and Technology.

[16] KITAGAWA, TOSIO (1959). Successive process of statistical inference applied to linear regression analysis and its specialization to response surface analyses. *Bull. Math. Statist.* **8** 80–114.

[17] LARSON, H. J. and BANCROFT, T. A. (1963). Biases in prediction by regression for certain incompletely specified models. *Biometrika* **50** 391–402.

[18] LARSON, H. J. and BANCROFT, T. A. (1963). Sequential model building for prediction in regression, I. *Ann. Math. Statist.* **34** 462–479.

[19] MEAD, R. J. (1968). Size and power of analysis of variance test procedures for incompletely specified fixed models. M.S. thesis, Iowa State Univ. of Science and Technology.

[20] PAULL, A. E. (1950). On a preliminary test for pooling mean squares in the analysis of variance. *Ann. Math. Statist.* **21** 539–556.

[21] SHATZOFF, M., FEINBERG, S. and TSAO, R. (1968). Efficient calculation of all possible regressions. *Technometrics* **10** 769–780.

[22] SNEDECOR, G. and COCHRAN, W. G. (1967). Statistical Methods (6th ed.). Iowa State Univ. Press.