DECUS
PROGRAM LIBRARY

DECUS NO.	FOCAL8-76
TITLE	SCREENING REGRESSION
AUTHOR	Dr. Robert G. Miller Glastonbury, Connecicut
COMPANY	
DATE	October 3, 1969
SOURCE LANGUAGE	FOCAL

SCREENING REGRESSION

DECUS Program Library Write-up — DECUS No. FOCAL8-76

OPERATING INSTRUCTIONS

1) A PDP-8 , 8/S , 8/I or 8/L with 4096 words and an ASR33 teletype.

2) FOCAL 1969 with no extended functions and the following suppressions or changes:

IN	FIND	PUT
1217	4551	7600
6002	4551	7600
2163	4551	7000
63	2676	1354
64	2666	2414
2732	6001	5336
2762	6046	7000

3) After reading in the FOCAL tape of "Screening Regression" start with GOTO 2.10.

4) Type in the correct answers to:

- P (the number of independent variables in the problem)
- N (the sample size over which the sums, sums of squares and sums of crossproducts were derived)
- SY (the sum of Y, i.e., ΣY)
- SSY (the sum of squares of Y, i.e., ΣY^2)
- F [the tabled F distribution value satisfying the following:]

  $F_{\alpha}(1, N)$ such that $= \frac{1}{20 \cdot P}$

5) Place in the reader a punched tape containing the following input with P independent variables (P=1, . . . , no limit).

$$\Sigma X_1,\ \Sigma X_1^2,\ \Sigma X_1X_2,\ \Sigma X_1X_3,\ \ldots\ldots,\ \Sigma X_1X_p,\ \Sigma X_1Y$$

$$\Sigma X_2,\ \Sigma X_2X_1,\ \Sigma X_2^2,\ \Sigma X_2X_3,\ \ldots\ldots,\ \Sigma X_2X_p,\ \Sigma X_2Y$$

$$\Sigma X_3,\ \Sigma X_3X_1,\ \Sigma X_3X_2,\ \Sigma X_3^2,\ \Sigma X_3X_4,\ \ldots\ldots,\ \Sigma X_3X_p,\ \Sigma X_3Y$$

$$\vdots$$

$$\Sigma X_p,\ \Sigma X_pX_1,\ \Sigma X_pX_2,\ \ldots\ldots,\ \Sigma X_pX_{p-1},\ \Sigma X_p^2,\ \Sigma X_p^2,\ \Sigma X_pY$$

6) Turn teletype reader to ON and wait for print of PASS 1's selected variable ID with its computed F value. If no variable is significant only E(O) will print out.

7) To run subsequent passes reenter the punched tape in the reader at the beginning. Continue doing this until the program stops selecting. This is determined when the next variable to be selected is not statistically significant or when the maximum of six variables has been achieved. The information on the seventh variable will be printed but will not be included in the regression equation.

8) The program will calculate and print the regression equation as follows:

B(O) XXX.XXXXX (Additive constant)

B(ID $X^{(1)}$) XXX.XXXXX (Coefficient for the first selected variable $X^{(1)}$)

B(ID $X^{(2)}$) XXX.XXXXX (Coefficient for the second selected variable $X^{(2)}$)

.
.
.

B(ID $X^{(S)}$) XXX.XXXXX (Coefficient for the Sth selected variable $X^{(S)}$)

C-FOCAL,1969

```
Ø1.5Ø  F   I=1-P,Ø;F J=1-P,I;S A(-J-I*P-1)=AC(-J-I*P)
Ø1.55  S   B=Ø;F Q=1,N;D 5
Ø1.7Ø  S   D(P)=Y;S R(P)=M;S X=X-B;F J=Ø,P;S A(P*J+J+P)=V(J)
Ø1.8Ø  I   (B*A/X-W)1Ø.1;T %3,P,D(P),%8.Ø2,B*A/X,!;I (6-P)1Ø.1;S P=P+1;G

Ø2.1Ø  T   "SCREENING REGRESSION",!
Ø2.2Ø  A   "P",N,!,"N",A,!,S Y",R,!,"SS Y",X,!, "F",W,!
Ø2.3Ø  T   "PASS SELVAR F",!;S P=1;S X=X-R↑2/A;G

Ø3.7Ø  F   J=1,P;S I=P;F L=Ø,J-1;D 9
Ø3.85  A   R(P);S J=P;S I=P;F L=Ø,J-1;D 6

Ø4.1Ø  S   Q=Q-1;G 1.7

Ø5.Ø5  F   J=Ø,P-1;D 7
Ø5.Ø7  S   D(P)=Q;I (N-Q)4.1;F L=Ø,N;A T;F J=Ø,P;D 8
Ø5.3Ø  D   3;S T=R(P)↑2/A(P↑2+2*P);I (B-T)5.5;R
Ø5.5Ø  S   B=T;S Y=Q;S M=R(P);F J=Ø,P;S V(J)=A(P*J+J+P)

Ø6.2Ø  S   T=J-1-L;S T=R(T)*A(P*T+T+J)/A(P*T+2*T);S R(P)=R(P)-T

Ø7.1Ø  I   (Q-D(J))7.2,7.3
Ø7.2Ø  R
Ø7.3Ø  S   Q=Q+1;F L=-1,N;A T

Ø8.1Ø  I   (L-D(J))8.2,8.3
Ø8.2Ø  R
Ø8.3Ø  S   A(J+P+J*P)=T

Ø9.2Ø  S   T=J-1-L;S T=A(I+T*P+T)*A(P*T+T+J)/A(T*P+2*T)
Ø9.3Ø  S   A(I+P*J+J)=A(I+P*J+J)-T

1Ø.1Ø  F   I=Ø,P;S A(P+I*(P+1))=R(I)
1Ø.2Ø  F   J=Ø,P-1;F K=J-P,Ø;D 12
1Ø.3Ø  S   V(P-1)=A((P+1)↑2-P-2);F J=2-P,Ø;D 11
1Ø.4Ø  F   I=Ø,P-1;T %2,"B",D(I),"  ",%6.Ø5,V(I),!
1Ø.5Ø  Q

11.1Ø  S   V(-J)=A((P+1)↑2-1-(P+1)*(P+J))
11.2Ø  S   T=Ø;F K=1,P+J-1;D 13
11.3Ø  S   V(-J)=V(-J)+T

12.1Ø  S   A(J+(P+1)*J-K)=A(J+(P+1)*J-K/A(J+(P+1)*J)

13.1Ø  S   T=T-A(-J+(P+1)*(-J)+K)*V(-J+K)
*
```

```
  *SCREENING REGRESSION
P
N
S Y
SS Y
F
PASS  SELVAR  F
    1     6     146.33
    2     3      21.95
    3     2      23.26
B   Ø     Ø.ØØØ16
B   6     Ø.14353
B   3     Ø.1Ø264
B   2     Ø.ØØØ61
*
```

REPRINTED FROM

OCTOBER 1962 METEOROLOGICAL MONOGRAPHS 45

APPENDIX A

A PROCEDURE FOR SELECTING PREDICTORS IN MULTIPLE REGRESSION ANALYSIS

In multiple regression a predictand, Y, is expressed as a linear function of a number of predictors, X_p ($p = 1, \cdots, P$),

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_pX_p + \cdots + a_PX_P \quad \text{(A-1)}$$

where the coefficients a_p ($p = 0, \cdots, P$) are determined using the method of least squares. The usual procedure for determining which of the P predictors are significant is to perform tests on the coefficients. If one or more predictors are not statistically significant, they may be eliminated from the equation by a method developed by Cochran [10]. The elimination of predictors is laborious if the original set of predictors P is large or if there are numerous nonsignificant predictors.

Experimenters seem to agree that in the application of multiple regression analysis to problems of prediction most of whatever predictability resides in an entire set of possible predictors can be found in a small subset of these predictors. For problems in which such is the case, it may be advantageous to attempt to select the subset of contributing predictors, leaving the redundant or noncontributing predictors out of the analysis. A method given by Bryan [7] attempts to perform just such a function. A description of the method will now be presented using the notation of Chapter I.

In order to select the first predictor $X^{(1)}$ the following quantities must be computed:

$$SST(Y) = \sum_{i=1}^{N} (Y_i - \bar{Y}_\cdot)^2$$

$$SST(X_p) = \sum_{i=1}^{N} (X_{pi} - \bar{X}_{p\cdot})^2 \quad \text{(A-2)}$$

$$SPT(YX_p) = \sum_{i=1}^{N} (Y_i - \bar{Y}_\cdot)(X_{pi} - \bar{X}_{p\cdot}) \quad (p = 1, \cdots, P)$$

where the number of observations in the dependent sample is N. Selection of the first predictor $X^{(1)}$, from the set of P possible predictors, is determined using the criterion[1]

$$\frac{SSF(X^{(1)})}{SSR(X^{(1)})} \geq \frac{SSF(X_p)}{SSR(X_p)} \quad (p = 1, \cdots, P) \quad \text{(A-3)}$$

where $SSF(X_p)$, the fitted sum of squares for predictor X_p on Y, is

$$SSF(X_p) = SPT(YX_p) \cdot [SST(X_p)]^{-1} \cdot SPT(YX_p). \quad \text{(A-4)}$$

and $SSR(X_p)$, the residual sum of squares after fitting Y with predictor X_p, is got by subtracting $SSF(X_p)$ from $SST(Y)$, the total sum of squares of Y, or

$$SSR(X_p) = SST(Y) - SSF(X_p). \quad \text{(A-5)}$$

The significance of $X^{(1)}$ may not be tested by the usual F ratio since the criterion is a function of the test statistic. An approximate test of significance for $X^{(1)}$ has been suggested [23] which uses a critical F value whose probability level is a function of the number of possible predictors P. This critical F value, $F_{\alpha^*}{}^*$, is defined as

$$F_{\alpha^*}{}^* \equiv F_\alpha \quad \text{(A-6)}$$

where α^* is the size desired in the selection test, and α is the corresponding size of the tabled F distribution. For a fixed size α^*, the value of α is taken to be approximately equal to α^*/P. The critical value of $F_{\alpha^*}{}^*$ for testing the significance of the Sth selected predictor $X^{(S)}$ becomes

$$F_{\alpha^*}{}^* = F_{\alpha^*/(P-S+1)}. \quad \text{(A-7)}$$

Therefore, the predictor $X^{(1)}$ is considered significant if

$$(N-2)\frac{SSF(X^{(1)})}{SSR(X^{(1)})} > F_{(\alpha^*/P)}(1, N-2). \quad \text{(A-8)}$$

Should $X^{(1)}$ be significant it then becomes the first of r predictors to be selected from the original set of P possible predictors. If $X^{(1)}$ is not significant, no predictors are selected.

[1] In the rare event that two or more of the P predictors satisfy the inequality in (A-3), the procedure is to select, arbitrarily, $X^{(1)}$ to be the first of these in the order $1, \cdots, P$. This procedure is to be followed for selection of subsequent predictors when more than one predictor satisfies the selection criterion. One further point, quotients are actually not necessary in (A-3) for determining $X^{(1)}$. It would be sufficient computationally to use the criterion $SSF(X^{(1)}) \geq SSF(X_p)$. The quotient form is given to show the parallel between regression and discriminant analysis selection.

The choice of the second selected predictor $X^{(2)}$ requires the quantities

$$SPT(X^{(1)}X_p) = \sum_{i=1}^{N} (X_i^{(1)} - \bar{X}_\cdot^{(1)})(X_{pi} - \bar{X}_{p\cdot}) \quad (p = 1, \cdots, P; X_p \neq X^{(1)}). \tag{A-9}$$

The criterion for choosing $X^{(2)}$ is

$$\frac{SSF(X^{(2)}|X^{(1)})}{SSR(X^{(2)}|X^{(1)})} \geq \frac{SSF(X_p|X^{(1)})}{SSR(X_p|X^{(1)})} \quad (p = 1, \cdots, P; X_p \neq X^{(1)}) \tag{A-10}$$

where

$$SSF(X_p|X^{(1)}) = \begin{bmatrix} SPT(YX^{(1)}) \\ SPT(YX_p) \end{bmatrix}' \begin{bmatrix} SST(X^{(1)}) & SPT(X^{(1)}X_p) \\ SPT(X^{(1)}X_p) & SST(X_p) \end{bmatrix}^{-1} \begin{bmatrix} SPT(YX^{(1)}) \\ SPT(YX_p) \end{bmatrix} - SSF(X^{(1)})$$
$$(p = 1, \cdots, P; X_p \neq X^{(1)}) \tag{A-11}$$

and

$$SSR(X_p|X^{(1)}) = SST(Y) - SSF(X^{(1)}) - SSF(X_p|X^{(1)}) \quad (p = 1, \cdots, P). \tag{A-12}$$

Predictor $X^{(2)}$ is judged significant if, from (A-7),

$$(N - 3)\frac{SSF(X^{(2)}|X^{(1)})}{SSR(X^{(2)}|X^{(1)})} > F_{\alpha^*/(P-1)}(1, N - 3). \tag{A-13}$$

Selection of predictor $X^{(S)}$ is made using the following general form:

$$\frac{SSF(X^{(S)}|X^{(1)}\cdots X^{(S-1)})}{SSR(X^{(S)}|X^{(1)}\cdots X^{(S-1)})} \geq \frac{SSF(X_p|X^{(1)}\cdots X^{(S-1)})}{SSR(X_p|X^{(1)}\cdots X^{(S-1)})} \quad (p = 1, \cdots, P; X_p \neq X^{(1)}\cdots X^{(S-1)}) \tag{A-14}$$

where

$$SSF(X_p|X^{(1)}\cdots X^{(S-1)}) = \begin{bmatrix} SPT(YX^{(1)}) \\ \vdots \\ SPT(YX^{(S-1)}) \\ SPT(YX_p) \end{bmatrix}' \begin{bmatrix} SST(X^{(1)}) & \cdots & SPT(X^{(1)}X^{(S-1)}) & SPT(X^{(1)}X_p) \\ \vdots & & \vdots & \vdots \\ SPT(X^{(1)}X^{(S-1)}) & \cdots & SST(X^{(S-1)}) & SPT(X^{(S-1)}X_p) \\ SPT(X^{(1)}X_p) & \cdots & SPT(X^{(S-1)}X_p) & SST(X_p) \end{bmatrix}^{-1}$$
$$\cdot \begin{bmatrix} SPT(YX^{(1)}) \\ \vdots \\ SPT(YX^{(S-1)}) \\ SPT(YX_p) \end{bmatrix} - SSF(X^{(1)}) - \cdots - SSF(X^{(S-1)}|X^{(1)}\cdots X^{(S-2)})$$
$$(p = 1, \cdots, P; X_p \neq X^{(1)}\cdots X^{(S-1)}) \tag{A-15}$$

and

$$SSR(X_p|X^{(1)}\cdots X^{(S-1)}) = SST(Y) - SSF(X^{(1)}) - SSF(X^{(2)}|X^{(1)}) - \cdots - SSF(X_p|X^{(1)}\cdots X^{(S-1)})$$
$$(p = 1, \cdots, P; X_p \neq X^{(1)}\cdots X^{(S-1)}). \tag{A-16}$$

Predictor $X^{(S)}$ is said to be significant if

$$[N - (S + 1)]\frac{SSF(X^{(S)}|X^{(1)}\cdots X^{(S-1)})}{SSR(X^{(S)}|X^{(1)}\cdots X^{(S+1)})} > F_{\alpha^*/(P-S+1)}(1, N - (S + 1)). \tag{A-17}$$

For a particular regression problem the number of selected predictors r is determined such that the predictor $X^{(r+1)}$ fails to show significance.

This procedure cannot be said to select necessarily the best set of r predictors out of the original set containing P predictors. However, it has been shown that it can select a highly reliable set of predictors when applied to particular problems in meteorology [24; 27; 35].

This method may leave a set of unselected predictors which together contain significant predictive information. Because none of these alone gives a significant contribution they all remain unselected. This can be remedied in part by testing significance on combinations of selected variables. In practice, however, it has been found that F^* tends to work well as a significance test only when variables are considered singly. This may be a consequence of the fact that a bias is introduced in the regression coefficients as a

result of selection. A critical evaluation of the amount of bias created by selection has not as yet been carried out.

The method contains the labor-saving feature of not requiring the crossproducts between unselected predictors. When all P variables must be processed in ordinary multiple regression $\frac{P(P+1)}{2}$ sums of squares or crossproducts are required. In the selection procedure only $P \cdot r - \frac{r(r-1)}{2}$ are necessary. If P is large this saving is sizable