DOCUMENT RESUME

ED 171 746 TM 008 824

AUTHOR	Pohlmann, John T.
TITLE	Controlling the Type I Error Rate in Stepwise Regression Analysis.
PUB DATE	Apr 79
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979); appendices marginally legible
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Computer Programs; *Error Patterns; *Mathematical Models; *Multiple Regression Analysis; Predictor Variables; Research Design; Statistical Analysis; *Statistical Studies; *Tests of Significance
IDENTIFIERS	*Stepwise Regression; *Type I Errors

ABSTRACT

Three procedures used to control Type I error rate in stepwise regression analysis are forward selection, backward elimination, and true stepwise. In the forward selection method, a model of the dependent variable is formed by choosing the single best predictor; then the second predictor which makes the strongest contribution to the prediction of the dependent variable is chosen, controlling for the effects of the first variable. The process continues so that the variable chosen increases the prediction potential, until remaining variables fail to make any contribution. Backward elimination begins with a model containing all predictors; and, at each step, a variable is eliminated if its removal results in the smallest reduction of effectiveness. True stepwise procedure is a variant of forward selection. To test these procedures, a Monte Carlo computer program, written in FORTRAN IV, was prepared. The results support two conclusions: (1) the probability of erroneously forming a regression model increases as a function of the number of predictors; and (2) as the inter-predictor correlation increases, the probability of making errors decreases. Therefore, the number of predictors and the inter-predictor correlation should be considered when attempting to solve an error rate problem. (MH)

Controlling the Type I Error Rate in Stepwise Regression Analysis

John T. Pohlmann

Southern Illinois University, Carbondale

AERA

April, 1979

San Francisco

Session 24.19

Controlling the Type I Error Rate

in Stepwise Regression Analysis

Stepwise regression has become a widely used technique for selecting a subset of potential predictors for some dependent variable. Three procedures have been used under the rubric of stepwise regression analysis: Forward selection, backward elimination, and true stepwise (Draper and Smith, 19

The forward selection procedure forms a model of the dependent variable by first selecting the best single predictor, then the second predictor is chosen which makes the strongest contribution to the prediction of Y, controlling for the effect of the first predictor. The process continues so that at each step, the variable selected for inclusion in the model increases the prediction of Y more than any other predictor. The selection process stops when the remaining variables fail to contribute significantly to the prediction of Y. The backward elimination procedure begins with a model containing all potential predictors, and then at each step a variable is eliminated if its removal from the model results in the smallest reduction in the model's effectiveness. The elimination process continues until the removal of any variable results in a significant reduction in the model's R^2. The true stepwise procedure is a variant of the forward selection technique. It differs from the forward selection procedure in that at each step, a variable that has been previously included in the model may be deleted if a partial F-test shows that variable to be an insignificant predictor.

In most of the computer statistical packages that have stepwise regression procedures, the criterion used for variable selection is an F-test formed as follows:

$$F = \frac{R_F^2 - R_R^2}{(1-R_F^2)/(N-p-1)} \tag{1}$$

where:

R_F^2 = the coefficient of determination for the model containing all predictors included at previous steps plus the variable under test.

R_R^2 = the coefficient of determination for the model containing all predictors except the variable under test.

N = the number of observations.

p = the number of predictors used in the model that produced R_F^2.

As with any statistical test, two kinds of inferential errors can be made. A type I error would occur if a variable was selected, using the F ratio criterion, when that variable's population regression weight was zero. A type II error occurs when a variable is not selected, using the F-test criterion, when that variable has a non-zero population weight.

Most users of stepwise regression adopt one of the traditional significance levels (.05 or .01) when evaluating the F-test in (1). This significance level will determine the type I error rate for each test. However, another perspective can be taken when considering the type I error rate, the problem-wide error rate.

The problem-wide error rate is the probability of selecting any variable when all variables have population regression weights of zero. In other words, the problem-wide error rate is the probability of forming a sample regression model, when none should be formed. The rest of this paper addresses this error rate, and a procedure will be presented that allows researchers to control its value.

The problem-wide error rate is comparable to the family-wide error rate commonly encountered in the context of <u>post hoc</u> tests conducted after a significant effect has been found in an ANOVA. For example, the probability of making one or more type I errors in a family of orthogonal tests is:

$$\alpha_F = 1 - \prod_{i=1}^{k} (1-\alpha_i) \qquad (2)$$

where α_F = the family-wide error rate.

k = the number of orthogonal tests.

α_i = the significance level on test i.

When the α_i's are all equal to α_T,

$$\alpha_F = 1 - (1-\alpha_T)^k. \qquad (3)$$

If a researcher wished to control α_F by reducing α_T, (3) could be solved for α_T:

$$\alpha_T = 1 - \sqrt[k]{1-\alpha_F} . \tag{4}$$

Alternately, the researcher could conservatively approximate α_T using the Bonnferoni inequality,

$$\alpha_T \approx \alpha_F/k \tag{5}$$

When the members of the family of tests are not orthogonal, formulas (4) and (5) yield conservative values of α_T. That is, the use of α_T from (4) or (5) will result in an α_F less than the desired value. The solution for α_T is considerably more complex when the tests are not orthogonal. The solution for a critical F that will maintain α_F at a desired value should be done using the correlated F distributon (Pope and Webster, 1972). Unfortunately the integration of the correlated F distribution is an extremely tedious process, and only limited tables critical values derived from it are available. Consequently, an approximate solution was sought using Monte Carlo methods.

METHOD

A Monte Carlo program, written in FORTRAN IV, was prepared by the author for this project. The program incorporated subroutines supplied in the International Mathematical and Statistical Library (1975). The IMSL subroutines were selected because of their proven accuracy and efficiency. A copy of the program is supplied in the Appendix of this paper.

The program generates sample data matrices (cases by variables) sampled with a given population dispersion matrix. Subroutine GGNRM was used for this purpose. Various population correlation matrices were supplied to GGNRM and a sample data matrix of standard normal deviates was produced. All population correlations between the predictors and the criterion variable were set equal to zero. The inter-predictor correlations were all set equal to a common value, and for the various replications examined in this study, the inter-predictor correlations were 0, .3, .5, . . and .9. In addition, the numbers of predictors used were 1, 3, [illegible], [illegible], [illegible], and 20. For every combination of the number of predictors and the average inter-predictor correlation (35 in all), a thousand sample data sets were generated.

Each data set thus generated was then subjected to a stepwise regression analysis using IMSL subroutine RLSTEP. Subroutine RLSTEP uses a true stepwise procedure. Variable selection is governed by a significance testing process. When, at any step, no F-test is significant, the selection process ceases.

For the purposes of this study, an error occured when a model, other than the null model, was formed by subroutine RLSTEP. The proportion of analyses resulting in a model was treated as an empirical estimate of the probability of erroneously forming a model using stepwise regression analysis.

RESULTS

Table 1 shows the results obtained when a variable selection significance level of .05 is used. The table entries in Table 1 are the proportion of 1000 stepwise regression analyses that produced a sample model when none should have been produced. For example, when a researcher has ten potential predictors that have correlations with each other equal to .50, the probability of erroneously forming a model is approximately .308.

Since the values in Table 1 are empirical estimates of the actual probabilities of making an error, there is some sampling error. The magnitude of the sampling error can be conservatively estimated by using the standard error of a proportion when p = .5. Since 1000 replications were used to derive each table entry, the standard error of a sample proportion will be less than or equal to .016. Consequently, a conservative 68% confidence interval for the true probability of making an error will be tabled value ± .016.

The figures in Table 1 support two conclusions: (1) The probability of erroneously forming a regression model increases dramatically as a function of the number of predictors, and (2), as the inter-predictor correlation increases, the probability of making an error decreases. Consequently, any solution to the error rate problem must take into consideration the number of predictors and the inter-predictor correlation.

After Table 1 was prepared, an attempt was made to develop an algorithm that could be used to select a significance level for variable selection that would control the problem-wide error rate.

The rationale for the algorithm presented here was based on the formula that gives the family-wide error rate in k independent tests. Formula (3) is reproduced here for this purpose:

$$\alpha_F = 1 - (1-\alpha_T)^k \,, \tag{6}$$

All terms are defined in (3). If α_T and α_F are known, k can be solved for as follows:

$$k = \frac{\ln(1-\alpha_F)}{\ln(1-\alpha_T)} \tag{7}$$

Formula (7) was applied to each entry in Table 1, and the resulting k values are given in Table 2. In producing Table 2, α_T was .05 and α_F was taken as the corresponding value in Table 1. The k values in Table 2 were then plotted as a function of various measures of the inter-predictor correlation. Figure 1 shows one of these plots for the 10 predictor variable case. The k values were observed to be an inverse linear function of ρ^2_{xx}, the inter-predictor correlation. The following function was considered to be a reasonable approximation:

$$k = p - (p-1)\rho^2_{xx} \tag{8}$$

where p = the number of predictors

ρ^2_{xx} = the inter-predictor correlation.

This function seemed suitable since for the extreme cases of ρ^2_{xx}, 0 and 1.0, (8) produced k values of p and 1 respectively. When ρ^2_{xx} is equal to 0, the problem-wide error rate should equal the α_F value given by (6). Under this condition ($\rho^2_{xx} = 0$) the error rate is directly analogous to the family-wide error rate for a family of orthogonal tests. When ρ^2_{xx} is equal to 1, every predictor is linearly dependent on the other predictors, hence there is in fact only one predictor. Formula (8) yields a k value of 1, when ρ^2_{xx} equals 1. In addition, inspection of plots, such as Figure 1, suggested that (8) was also accurate for estimating k for values of ρ^2_{xx} between 0 and 1.

Unfortunately a researcher using stepwise regression never knows ρ^2_{xx}, so it must be estimated. A less biased estimate of the squared correlation coefficient can be obtained using the shrinkage formula (McNumar, 1969):

$$\hat{\rho}^2 = 1 - (1-r^2)\frac{N-1}{N-2} \quad . \tag{9}$$

The estimate of ρ^2_{xx} used for this study was obtained as follows:

Let R_{pp} = the inter-predictor correlation matrix.

Define each element in $\hat{R}_{pp}$ as

$$\hat{r}^2_{ij} = 1 - (1-r^2_{ij})\,\frac{N-1}{N-2}\ , \tag{10}$$

where r^2_{ij} = the square of the ijth element of R_{pp}, and

N = the number of observations.

$$\text{Let } \bar{r}^2 = \frac{\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \hat{r}^2_{ij}}{\frac{1}{2}(p^2-p)}\ , \tag{11}$$

which is the mean of the off diagonal elements of $\hat{R}_{pp}$.

The sample estimate of ρ^2_{xx} is then substituted into (8) to obtain

$$k = p - (p-1)\bar{r}^2\ . \tag{12}$$

After k has been obtained via (12), α_T is obtained.

$$\alpha_T = 1 - \sqrt[k]{1-\alpha_F} \quad , \tag{13}$$

where α_F is the desired problem-wide error rate. A concise worked example is given in the Appendix of this paper.

The validity of the proposed algorithm was then tested by modifying the Monte Carlo program, used to produce Table 1, to use (13) to select an α_T. The results of this validation study are presented in Table 3. As can be noted in Table 3, the probability of erroneously forming a model, using (13) to determine α_T, approaches the desired value of .05. There is a slight tendency for this procedure to produce conservative values of α_T. The average value of α_F in Table 3 is .045, and the conservative nature of the procedure is most apparent for problems with large numbers of predictors and high inter-predictor correlations.

DISCUSSION

The type I error rate in stepwise regression analysis deserves serious consideration by researchers. The literature is replete with "significant" findings that fail the ultimate test of replication. One possible explanation for this state of affairs might lie in the increasing problem-wide error rate that can occur in stepwise regression analysis.

If a researcher considers the problem wide error rate important, he or she should take some corrective action. Three possibilities exist, depending on the kind of analysis contemplated. They are: (1) Prior to the stepwise analysis conduct an omnibus test of the model containing all potential predictors, (2) use the backward elimination procedure and use an α_T obtained by substituting the number of predictors for k in (13), or (3) use the algorithm for obtaining α_T presented here, if a forward selection or true stepwise procedure is used.

The Omnibus Test

The analysis begins by forming a full model containing all predictors. The R^2 for this model is tested for significance at the α_F level. The F is obtained as follows:

$$F = \frac{R^2/p}{(1-R^2)/(N-p-1)} \quad , \tag{14}$$

where R^2 = the coefficient of determination for the model containing all potential predictors,

p = the number of predictors,

N = the number of cases.

This F ratio yields a simultaneous test of significance for all weights in a model. Proceed with the analysis only if a significant F using (14) is obtained.

The Backward Elimination Procedure

The backward elimination procedure is comparable to testing a family of orthogonal hypotheses. At each step, the variance accounted for in the dependent variable that is tested for each predictor is independent of all ther sources of variation. Consequently, the use of

$$\alpha_T = 1 - \sqrt[p]{1-\alpha_F} \quad , \tag{15}$$

will maintain α_F at its desired value.

Finally, the algorithm developed in this paper is recommended if a forward selection or true stepwise procedure is used. Since the value of α_T obtained using (13) will be greater than that obtained using (15), when some covariance among the predictors is present, the use of (13) will produce a more powerful analysis.

References

Draper, N. R., & Smith, H. Applied regression analysis. New York: John Wiley & Sons, 1966.

International mathematical and statistical library (5th Ed.). Houston: International Mathematical and Statistical Libraries, Inc., 1975.

McNemar, Q. Psychological statistics. New York: John Wiley & Sons, 1969.

Pope, P. T., & Webster, J. T. The use of an F-statistic in stepwise regression procedures. Technometrics, 1972, 14(2), 327-340.

Table 1

Monte Carlo Estimates of the Probability of

Erroneously Forming a Sample Model Using

Stepwise Regression Analysis with

a Variable Selection Significance Level of .05

Inter-Predictor	Number of Predictors						
Correlation	2	3	4	5	7	10	20
.0	.102	.130	.184	.216	.304	.410	.653
.3	.101	.130	.178	.213	.275	.367	.552
.5	.097	.128	.171	.196	.235	.308	.417
.7	.085	.125	.140	.153	.185	.225	.314
.9	.073	.094	.101	.111	.122	.126	.169

Table 2

k Values Derived Using Formula (7)

on the Values from Table 1

Inter-Predictor Correlation	Number of Predictors						
	2	3	4	5	7	10	20
.0	2.10	2.72	3.96	4.74	7.06	10.29	20.63
.3	2.08	2.72	3.82	4.67	6.27	8.92	15.70
.5	1.99	2.67	3.66	4.25	5.22	7.18	10.52
.7	1.73	2.60	2.94	3.24	3.98	4.97	7.35
.9	1.48	1.92	2.08	2.28	2.54	2.63	3.61

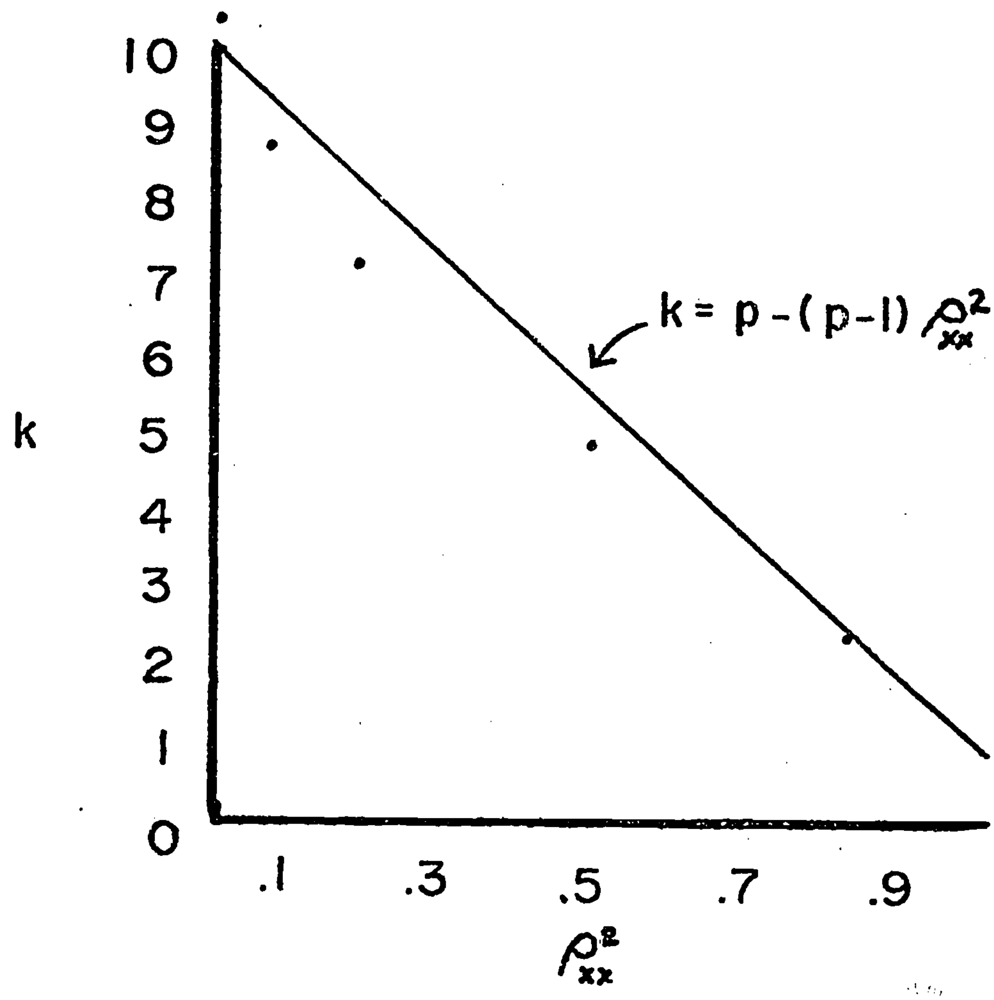

Figure 1. Plot of k as a function of ρ_{xx}^2 for 10 predictors.

Table 3

Monte Carlo Estimates of the Probability of Erroneously Forming a Sample Model Using Stepwise Regression Analysis with a Variable Selection Significance Level Obtained Using Formula 13. The Desired α_F was .05

Inter-Predictor	Number of Predictors						
Correlation	2	3	4	5	7	10	20
.0	.052	.044	.058	.044	.048	.045	.055
.3	.050	.045	.044	.055	.046	.047	.038
.5	.060	.044	.041	.063	.041	.044	.042
.7	.059	.050	.041	.046	.037	.031	.032
.9	.045	.054	.056	.050	.033	.027	.011

APPENDIX I

Copy of the Computer Program Used
in the Study

```
      REAL*8 DD
      COMMON/RLSTL/DD,IH(11)
      [illegible]
      WRITE(6,15)
15    FORMAT (' INPUT SEED')
      READ(5,*) ISEED
      WRITE(6,16)
16    FORMAT(1X,'INPUT [illegible] ALPHA [illegible]')
      READ(5,*,END=999)N,K,ALPHIN,ALPOUT
      WRITE(6,20)
20    FORMAT(' INPUT LOWER TRIANGLE OF POP [illegible] MATRIX [illegible]')
      L=(K*(K+1))/2
      READ(5,*,END=999)(SIG1(I),I=1,L)
      WRITE(6,501)
501   FORMAT(' INPUT NUMBER OF SAMPLES')
      READ(5,*)NSAMP
      TYPE1=0.
      DO 500 ISAMP=1,NSAMP
C     WRITE(6,502)ISAMP
502   FORMAT(1X,10('*'),'SAMPLE',I4,10('*'))
      DO 50 I=1,L
50    SIG2(I)=SIG1(I)
      CALL GAMMA(ISEED,N,K,SIG1,200,X,[illegible],IER)
      DO 40 I=1,L
40    SIG1(I)=SIG2(I)
C     DO 100 I=1,N
C100  WRITE(6,60)(X(I,J),J=1,K)
60    FORMAT(' ',10F7.2)
      CALL BECOVI (X,N,K,200,XM,S,R,IER)
C     WRITE(6,70)
70    FORMAT(' MEANS AND SDS')
C     DO 200 I=1,K
C200  WRITE(6,90)I,XM(I),S(I)
90    FORMAT(1X,I5,2F10.3)
C     WRITE(6,120)
120   FORMAT(' R MATRIX')
      L2=0
C     DO 250 I=1,K
C     L2=L2+I
C     L3=L2-I+1
C250  WRITE(6,220)(R(J),J=L3,L2)
      M=K-1
      DO 410 I=1,20
410   IH(I)=0
      IOPT=0
      CALL RLSTEP(R,M,N,ALPHIN,ALPOUT,IH,4,IOPT,IER)
      NP=0
      DO 420 I=1,4
      IF(IH(I).EQ.1)NP=NP+1
420   CONTINUE
      IF (NP.GT.0))TYPE1=TYPE1+1.0
500   CONTINUE
      TYPE1=TYPE1/FLOAT(NSAMP)

      WRITE(6,510)TYPE1
510   FORMAT(' PROPORTION OF SAMPLES FORMING A MODEL = ',F7.3)
220   FORMAT (1X,10F7.3)
      [illegible]
999   STOP
      END
```

APPENDIX II

A Worked Example of the Algorithm for Obtaining a Significance Level for Variable Selection Using Stepwise Regression

R_{pp} N = 20 Desired Model Error Rate = .05

$$\begin{bmatrix} 1.0 & .3 & .5 \\ & 1.0 & .2 \\ \text{sym} & & 1.0 \end{bmatrix}$$

$$\hat{r}^2_{ij} = 1 - (1 - r^2_{ij}) \frac{N - 1}{N - 2}$$

$$\hat{r}^2_{12} = .0394$$

$$\hat{r}^2_{13} = .2083$$

$$\hat{r}^2_{23} = -.0133$$

$$\bar{r}^2 = \frac{\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \hat{r}^2_{ij}}{\frac{1}{2}(p^2 - p)} = .0781$$

$$k = p - (p - 1)\,\bar{r}^2 = 2.8438$$

$$\alpha_T = 1 - \sqrt[k]{1 - \alpha_F} = .0179$$