ED 368 787 TM 021 304

AUTHOR	Schumacker, Randall E.
TITLE	A Comparison of Best Model Selection Criteria in Multiple Regression.
PUB DATE	Jan 94
NOTE	35p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 27-29, 1994).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Criteria; Models; Prediction; *Regression (Statistics); Research Problems; *Selection; *Statistical Studies
IDENTIFIERS	All Possible Subset Approach; Principal Components Analysis; Stepwise Regression; *Subset Analysis

ABSTRACT

The all-possible subset approach is recommended as an alternative over stepwise methods for selecting the best set of predictor variables for multiple regression. Several criteria are available for selecting the best subset model. These are compared with the principal component regression (PCR) method to investigate their usefulness for subset model selection. An applied example illustrates the comparisons. Subjects were 80 females and 76 males who participated in an early college admission program for gifted students entering the Texas Academy of Mathematics and Science. Their scores on the Learning and Study Strategies Inventory (LASSI) were analyzed; the research question was whether LASSI subscales could predict a student's college grade point average. The best subset model for each subset size with the corresponding selection criteria is presented in a table. Six tables and one figure present analysis results. An appendix contains the Statistical Analysis System program for the analysis. (Contains 22 references.) (SLD)

A Comparison of Best Model Selection Criteria in Multiple Regression

Randall E. Schumacker, Ph.D.
Educational Research
College of Education
University of North Texas
Denton, TX 76203-6857
(817) 565-3962
schmckr@coe.unt.edu

Paper presented at the annual meeting of the
Southwest Educational Research Association
January 27-29, 1994 San Antonio, Texas

A Comparison of Best Model Selection Criteria
in
Multiple Regression

Multiple regression permits model testing wherein a set of independent variables are hypothesized to predict a dependent variable. Oftentimes when the set of variables selected do not significantly predict, the researcher searches for a "subset" of variables that provides the best pred.ction model. The various multiple regression stepwise methods have been extensively used for this purpose. Prior research, however, has indicated that the all possible subset approach is preferred over the stepwise methods in determining the best model (Berk, 1978; Thayer, 1986; Davidson, 1988; Henderson & Denison, 1989; Welge, 1990; Thayer, 1990). Thompson et al. (1991), in further crticizing stepwise methods, recommended that effect sizes be computed for each "all possible" subset equation and that the subset model which has the desired effect size be chosen.

Zuccaro (1992) investigated the use of the C_p criteria in contrast to the stepwise methods for determining the best set of predictors using a sample data set. The C_p statistic measures the total squared error variance in each subset model containing **p** predictor variables [error variance plus the bias introduced by not including important variables]. Findings suggested that the selection of the best subset model with the lowest bias is indicated by the smallest Mallows' C_p criteria (Mallows, 1966; 1973), especially in the presence of multicollinearity. Pohlmann (1983) noted that multicollinearity among predictor variables didn't affect the Type I error rate, but did affect the Type II error rate and width of the confidence interval. Findings suggested that sample size and model validity could compensate for multicollinearity effects, especially when certain research questions required models with highly correlated predictors, e.g. $Y = \beta_1 X_1 + \beta_2 X^2_1 + e$.

The principal component regression (PCR) approach has also been proposed as a criteria for selecting the best predictor model. The method appears to be useful when predicting values in one sample based upon estimates from another sample and when multicollinearity exists among a set of variables (Morrison, 1976). The rationale for using a PCR approach is when the mean squared error of a biased estimate is smaller than the variance of an unbiased estimate. The PCR method, however, is not appropriate for multiple regression subset models containing interactions (Aiken & West, 1993) nor when models depict nonlinear correlated predicter sets. The PCR method creates a set of new variables called principal components, which are uncorrelated or orthogonal, and therefore preclude it from being used in these types of models.

A review of the literature indicated that researchers misuse stepwise methods to determine the best predictor set or interpret the importance of predictor variables (Huberty, 1989; Snyder, 1991; Thompson, 1989; Thompson et al., 1991, Welge, 1990). Stepwise methods inflate Type I error rates by not using the correct degrees of freedom in calculating the change in R^2.

Additionally, researchers incorrectly interpret the order of variable selection as defining the "best set" of variables in the predictor set. Also, the order of variable entry is misinterpreted as determining which variables are the important predictors.

The all possible subset approach is recommended as an alternative over stepwise methods for selecting the best set of predictor variables. Several criteria, however, are available for selecting the best subset model: R^2 , Adj. R^2, MSE, C_p, or the principal component regression method. How do these criteria compare when selecting the best subset model? When might a researcher choose one criteria over another for selecting the best model? The principal component regression method, which determines the best model for prediction by redefining the theoretical model, appears more useful when estimates from one sample are used to predict in another sample. The C_p statistic is useful when the predictor set is correlated, whereas the principal component method creates a set of orthogonal predictors. A comparison of the selection criteria and the PCR approach will permit an investigation of their usefulness for subset model selection. An applied example will illustrate a comparison of the criteria and afford further discussion. The objective of this study, therefore, was to compare the various model subset selection criteria and provide guidelines for the selection of the best "subset" model.

METHODS AND PROCEDURES

Subjects

Subjects came from a cohort of students accepted into the Texas Academy of Mathematics and Science (TAMS) at the University of North Texas in Fall, 1993. TAMS is an early college entrance program in which students earn approximately 60 hours of college credit by taking University of North Texas courses. Students enter TAMS at the beginning of their 11th year in high school. They live on campus in a special residence hall and take regular university courses in mathematics, science and the humanities. After two years, participants receive a special high school diploma and have amassed at least 60 hours of college credit. Each year approximately 200 high school sophomores, who have met the selection criteria and completed the 10th grade, are accepted into the Texas Academy of Mathematics and Science.

In the study year, TAMS accepted 204 students. Of these, 156 students attended an August orientation, which occurred a week prior to their first semester of college coursework, and completed the LASSI. There were 80 females and 76 males who participated in the study. The students who took the LASSI were similar in demographic background and academic ability as previous classes because of the academy's consistent admission requirements and pool of applicants. The participants SAT-M and SAT-V means and standard deviations, respectively, were: *M*=651, *SD*=57; and *M*=530, *SD*=75.

Instrument

The LASSI is an English language assessment tool designed to measure college students' use of learning and study strategies. It was designed to provide assessment and pre-post achievement measures for students participating in a learning strategies and study skills project. A high-school version is available, but it was not recommended for use with accelerated students in these programs (Eldredge, 1990). The LASSI can be administered in a group setting in approximately 30 minutes. The carbonless test format allows participants to score their own assessment and take a copy of the results with them from the testing session.

The LASSI's ten subscales focus on thoughts and behaviors related to successful learning. The ten subscales are (1) attitude; (2) motivation; (3) time management; (4) anxiety; (5) concentration; (6) information processing; (7) selecting the main ideas; (8) study aids; (9) self-testing; and (10) test strategies (for more details see, Weinstein, 1987). Reliability studies reported Cronbach alpha internal consistency values ranging from .70 to .86 and test-retest reliabilities from .70 to .85. Validity studies have also reported normative data for high school and college students with different instruments for each group (Weinstein, Palmer, & Schulte, 1987). Students respond to individual items on each subscale using a five-point scale: (5) very typical of me; (4) fairly typical of me; (3) somewhat typical of me; (2) not very typical of me; and (1) not at all typical of me. Some item values are reverse keyed before being added to obtain a subscale score. The subscale scores are compared by graphing them onto a normal curve equivalent percentile chart.

According to the LASSI user's manual (Weinstein, 1987), students scoring above the 75th percentile do not need to improve that specific skill or strategy. Students scoring between the 75th percentile and the 50th percentile should consider improvement. Students scoring below the 50th percentile on a subscale need assistance to improve that skill or strategy. For example, students scoring below the 50th percentile on the anxiety subscale would be considered anxious about being in college. Likewise, students scoring below the 50th percentile on the motivation subscale lack appropriate motivation to do college level work effectively.

Research Question

The research question of interest was whether the ten LASSI subscales could predict a student's college grade point average after one semester of college coursework. A related question pertained to whether a "subset" of the ten LASSI subscales could better predict college grade point average for this sample of students. Students not maintaining at least a 2.50 grade point average after one semester of college coursework were dismissed from the Academy. Knowledge of which subscales are best predictors of college grade point average will aid staff in identifying potential at-risk students upon entering the Academy.

Data Analysis

The SAS statistical program is in the Appendix. The student's college grade point average was predicted by the ten LASSI subscales with the SELECTION statement requesting the best subset model criteria. A monotonic relationship existed between R^2 and Adjusted R^2, in fact, $r = 1.00$ across all subset sizes. A monotonic relationship also existed between C_p and MSE, in fact, $r = 1.00$ across all subset sizes. In addition, both R^2 and Adjusted R^2 had a perfect inverse relationship to both C_p and MSE, i.e. $r = -1.00$, across all subset sizes.

The R^2 statistic is calculated as the $SS_{regression}/SS_{total}$ and is the most commonly used and reporteu value in determining the proportion of variance in the dependent variable accounted for by the independent variables (Pedhazur, 1982). The adjusted R^2 value, which corrects for the number of predictors in the model, is computed as: $R^2 - [p(1-R^2)/N-p-1]$ (Norusis, 1979). In determining the number of variables to include in the regression model, the researcher will typically test for significant increments in R^2 values between models with differing numbers of predictors. The mean square error (MSE) value is an unbiased estimate of σ^2, the variance of ε, which represents random error and accounts for variation due to other factors. The MSE statistic is computed as: SS_{error}/df_2.

The Mallows' C_p statistic with the intercept term is calculated as: $C_p = (1-R^2_p)(n-T)/(1-R^2_T)-(n - 2p)$, or alternatively without the intercept term as: $C_p = (SSE_p/MSE) - (n - 2p) + 1$

(Freund & Littell, 1991). Mallows' C_p is useful in measuring the level of bias in the parameter estimates (β_j). The C_p criteria has also been recommended for determining the best set of predictors. The PROC PRINCOMP procedure was used to create ten orthogonal principal component variables. The principal component variable parameter estimates were computed using the PROC REG procedure. The number of significant principal component parameter estimates were then identified. These procedures are outlined in the *SAS System for Regression* manual (Freund & Littell, 1991).

RESULTS

The optimum subset model should generally be one that produces the minimum error sum of squares (MSE), or equivalently maximizes the R^2 value. The procedure for finding the optimum subset of all possible subset sizes requires computing 2^m equations. The ten subscale predictors in the model yielded 1024 regression equations (2^{10}) with associated selection criteria statistics {Note: The determination of the number of subset equations generated for p predictor variables from an m variable full model is: m!/[p!(m-p)!]. For example, the number of 2 variable subset equations (p) generated from a 10 variable model (m) would be 45}.

The correlation matrix, means and standard deviations of the ten LASSI subscales are in Table 1. The intercorrelations among the subscales indicated that Anxiety/Worry was not significantly correlated with Time Management, Information Processing, Support Techniques/Materials, and Self Testing/Class Preparation. The lowest subscale mean was on Selecting Main Ideas.

Insert Table 1 Here

The best subset model for each subset size with the corresponding selection criteria are in Table 2. A combined criteria, minimum error sum of squares (MSE) with maximum R^2, indicated a five variable subset model. In contrast, the C_p criteria indicated a four variable model. The four variable subset model for predicting college grade point average consisted of the four subscales: Motivation, Anxiety/Worry, Support Techniques/Materials, and Self Testing/Class Preparation. The fifth variable indicated in the combined criteria selection was Information Processing.

Insert Table 2 Here

The C_p criteria also indicated the bias in having too many variables in the model. Large C_p values indicated equations with larger mean square error. If $C_p > (p + 1)$, for any subset size p, then bias was present. If $C_p < (p + 1)$, for any subset size p, then the model contained too many variables. A plot of the C_p

values against the number of predictors, compared to a plot of the (p + 1) values, has been recommended for determining the best subset model (Mallows, 1973).

The present pattern of C_p values for the various subsets of size p are typical when multicollinearity is present. The C_p values initially become smaller, but then start to increase. The plot of C_p values is similar to a "scree" plot in factor analysis and as such a multiple regression method might also be useful in determining the number of variables to retain (Zoski & Jurs, 1993). The best subset model is indicated when the C_p values begin to increase and cross the (p + 1) values (see Figure 1).

Insert Figure 1 Here

Principal components were obtained by computing eigenvalues from the correlation matrix. The correlation matrix was used so that variables were not affected by the scale of measurement as in the use of a variance-covariance matrix. Since eigenvalues are the variances of the principal component variables, the sum of the eigenvalues equal the number of variables in the full model, just as the sum of standardized variable variances would equal the number of variables. This sum is the measure of the total variation in the data set.

A wide variation in the eigenvalues would sugge: : the presence of multicollinearity among the variables. The number of eigenvalues greater than unity, as in factor analysis, would indicate the number of variables from the full model that would explain most of the variance in the data set. The eigenvectors, in contrast, contain the coefficients for each principal component variable. These coefficients are used to create the observed values of the original variables. These observed values are then used in multiple regression as orthogonal, uncorrelated predictor values with no multicollinearity present. Table 3 contains the eigenvalues for the ten principal component variables generated from the correlation matrix of the ten LASSI subcales.

Insert Table 3 Here

Preliminary inspection of the eigenvalues indicates three principal component variables that account for 72.4 % of the variance in predicting college grade point average (7.24/10.00). The first pri cipal component alone explained 46 %. The ten principal component variables when analyzed in multiple regression yielded an $R^2 = .19$, Adj. $R^2 = .13$, and a MSE = .32 which is identical to the ten predictor model using the original variables obtained from the subset model approach (Table 4).

Insert Table 4 Here

A summary of parameter estimates in Table 5 indicates that model components 1, 4, and 8 are significant relative to other principal components in the full model. An examination of the coefficients in the eigenvectors for these principal components reveals which subscales contribute the most to the prediction of college grade point average (see Table 6).

Insert Table 5 Here

Insert Table 6 Here

The first principal component indicates that all subscales contributed to prediction. The fourth principal component indicates that Attention, Anxiety/Worry, and Information Processing are important. The eighth principal component comprised Attention, Motivation, Time Management, Concentration, Information Processing, and Support Techniques/Materials. The fourth and eighth principal components suggested "factors" which are secondary related to the primary construct tapped by the LASSI subscales. Providing names for the principal components, as in factor analysis, is subjective and only meaningful within the context of interpreting scores. These principal component results clearly indicate that all ten subscales, when treated as uncorrelated or orthogonal predictors, contributed to the prediction of future college grade point average the same as the set of ten correlated predictors in the ordinary least squares approach.

SUMMARY

The C_p criteria was the lowest for a four variable predictor model. This four variable subset model was verified by examining where the plot of C_p values against the (p + 1) values crossed. The MSE criteria was also the lowest for the five variable model. The R^2 and Adjusted R^2 criteria also indicated a five variable model. The full model with all ten subscales as predictors yielded the same result as the principal components method using the first principal component extracted. The first principal component yields the most variance accounted for in the set of variables. The C_p criteria selected the smallest variable subset model in the presence multicollinearity.

In using multiple regression it is important to have a theoretical basis for the regression model and to consider sample to sample fluctuations in R^2. A common misconception in multiple regression is that the model with all the significant predictors included is the best model. This isn't always the case. The problem is that the b's and R^2 values are data dependent due to the least squares criterion being applied to a specific sample of data. A different sample will usually result in different parameter estimates and variance explained. Although the standard errors of the b's do provide the researcher with some indication of the amount of change expected from sample to sample, the fact remains that the estimates obtained from one sample may predict poorly when applied to a new set of sample data. The primary method to assess the change in R^2 or b's is to replicate the regression model using other sample data. Bootstrapping, jacknifing, and cross-validation methods have also become useful in indicating the variation in b's and R^2 values when estimates from one sample are applied to another sample.

The rationale behind a regression model is to estimate σ^2 (the true model's mean square error variance). Since σ^2 is not generally known, a researcher must estimate it from a knowledge of prior research ($\sigma^2 = \sigma^2_{y.x}$), obtain estimates from a model containing all theoretically relevant predictors, replicate the study, or use bootstrapping, jacknifing, and cross-validation methods. In this regard, effect size considerations, as recommended by Thompson (1991), become important to consider.

REFERENCES

Aiken, L.S. & West, S.G. (1993). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: SAGE Publications.

Cummings, Corenna, C. (1982, March). *Estimates of multiple correlation coefficient shrinkage.* Paper presented at the American Educational Research Association annual meeting. New York, NY.

Davidson, Betty, M. (1988, November). *The case against using stepwise research methods.* Paper presented at the Mid-South Educational Research Association annual meeting. Louisville, KY.

Eldredge, J.L.' (1990). Learning and study strategies inventory: a high school version (test review). *Journal of Reading, 34,* 146-149.

Freund, R.J. & Littell, R.C. (1991). SAS System for Regression (2nd Ed.) SAS Institute: Cary, NC.

Henderson, Douglas, A. & Denison, Daniel R. (1989). Stepwise regression in social and psychological research. *Psychological Reports*, *64*(1), 251-257.

Huberty, C.J. (1989). *Problems with stepwise methods--better alternatives.* In B. Thompson (Ed.), *Advances in Social Science Methodology*, *1*, 43-70. Greenwich, CT: JAI Press.

Mallows, C.P. (1966). *Choosing a subset regression.* Paper presented at the Joint Statistical Meetings. Los Angeles, CA.

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

Norusis, M.J. (1979). *SPSS Statistical Algorithms*. SPSS, Inc.: Chicago, IL.

Pedhazur, E.J. (1982). *Multiple Regression in Behavioral Research (2nd). Explanation and Prediction*. CBS College Publishing, Hold Rinehart & Winston: New York, NY.

Pohlmann, J. (1983, April). *A perspective on multicollinearity.* Paper presented at the American Educational Research Association annual meeting. Montreal, Canada.

Snyder, P. (1991). *Three reasons why stepwise regression methods should not be used by researchers*. In B. Thompson (Ed.), *Advances in Educational Research: Substantive findings, methodological developments, 1*, 99-105. Greenwich, CT: JAI Press.

Thayer, Jerome, D. (1986, April). *Testing different model building procedures using multiple regression*. Paper presented at the American Educational Research Association annual meeting. San Francisco, CA.

Thayer, Jerome, D. (1990, April). *Implementing variable selection techniques in regression*. Paper presented at the American Educational Research Association annual meeting. Boston, MA.

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21(4)*, 146-148.

Thompson, B.; Smith, Q.W., Miller, L.M., & Thomson, W.A.. (1991, January). *Stepwise methods lead to bad interpretations: better alternatives*. Paper presented at the Southwest Educational Research Association annual meeting. San Antonio, TX.

Weinstein, C.E. (1987). *LASSI user's manual*. Clearwater, FL: H&H Publishing Company, Inc.

Weinstein, , C.E., Palmer, D.R., Schulte, A.C. (1987). *Learning an. Study Strategies Inventory*. Florida: H&H Publishing.

Welge, Patricia (1990, January). *Three reasons why stepwise regression methods should not be used by researchers*. Paper presented at the Southwest Educational Research Association annual meeting. Austin, TX.

Zoski, K.K. & Jurs, S.G. (1993). Using multiple regression to determine the number of factors to retain in factor analysis. *Multiple Linear Regression Viewpoints, 20*(1), 5-9.

Zuccaro, Cataldo (1992). Mallows' C_p statistic and model selection in multiple linear regression. *Journal of the Market Research Society, 34*(2), 163-172.

Table 1. LASSI Subscale inter-correlations, means and standard deviations (n = 156).

LASSI Subscale	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Attention	1.00									
(2) Motivation	.59	1.00								
(3) Time Management	.39	.60	1.00							
(4) Anxiety/Worry	.32	.15	.09	1.00						
(5) Concentration	.57	.62	.62	.33	1.00					
(6) Information	.20	.15	.39	.03	.26	1.00				
(7) Select Ideas	.25	.36	.31	.37	.47	.30	1.00			
(8) Support	.24	.40	.47	.05	.38	.45	.40	1.00		
(9) Class Prep.	.38	.50	.63	.06	.55	56	.39	.64	1.00	
(10)Test Strategy	.54	.47	.33	.50	.66	.20	.60	.21	.34	1.00
Mean	34.33	33.12	24.91	28.38	28.56	28.94	18.32	26.03	27.36	31.46
SD	4.17	4.73	6.18	5.92	4.93	5.24	3.51	5.96	5.84	4.58

Table 2. Best Model Selection Criteria by Subset Size.

Subset Size	Variables in Subset Model	R^2	AdjR^2	C(p)	MSE
1	(2)	.09	.08	10.88	.34
2	(2),(8)	.11	.10	8.01	.33
3	(2),(6),(8)	.14	.13	5.16	.32
4	(2),(4),(8),(9)	.17	.14	2.72	.32
5	(2),(4),(6),(8),(9)	.18	.15	2.93	.31
6	(2),(4),(6),(7),(8),(9)	.18	.15	3.68	.31
7	(1),(2),(4),(6),(7),(8),(9)	.19	.15	5.10	.31
8	(1),(2),(4),(6),(7),(8),(9),(10)	.19	.14	7.05	.32
9	(1),(2),(3),(4),(5),(6),(8),(9),(10)	.19	.13	10.04	.32
10	(1),(2),(3),(4),(5),(6),(7),(8),(9),(10)	.19	.13	11.00	.32

Table 3. Eigenvalues from Correlation Matrix: Principal Components

Principal Component	Eigenvalue	Difference	Proportion
(1)	4.64	3.07	.46
(2)	1.57	.54	.16
(3)	1.03	.38	.10
(4)	.65	.12	.07
(5)	.53	.02	.05
(6)	.51	.17	.05
(7)	.34	.07	.03
(8)	.27	.02	.03
(9)	.25	.04	.03
(10)	.21	-	.02
Σ	10.00		1.00

Table 4. Principal Component Regression

Model	SS	df	MS	F	p	R^2
Regression	10.76	10	1.08	3.35	.001	.19
Model Components						
(1)	4.16	1				
(2)	.99	1				
(3)	1.13	1				
(4)	1.93	1				
(5)	.09	1				
(6)	.23	1				
(7)	.58	1				
(8)	1.33	1				
(9)	.29	1				
(10)	.03	1				
Error	46.58	145	.32			
Total	57.34	155				

Note: Adj. R^2 = .13

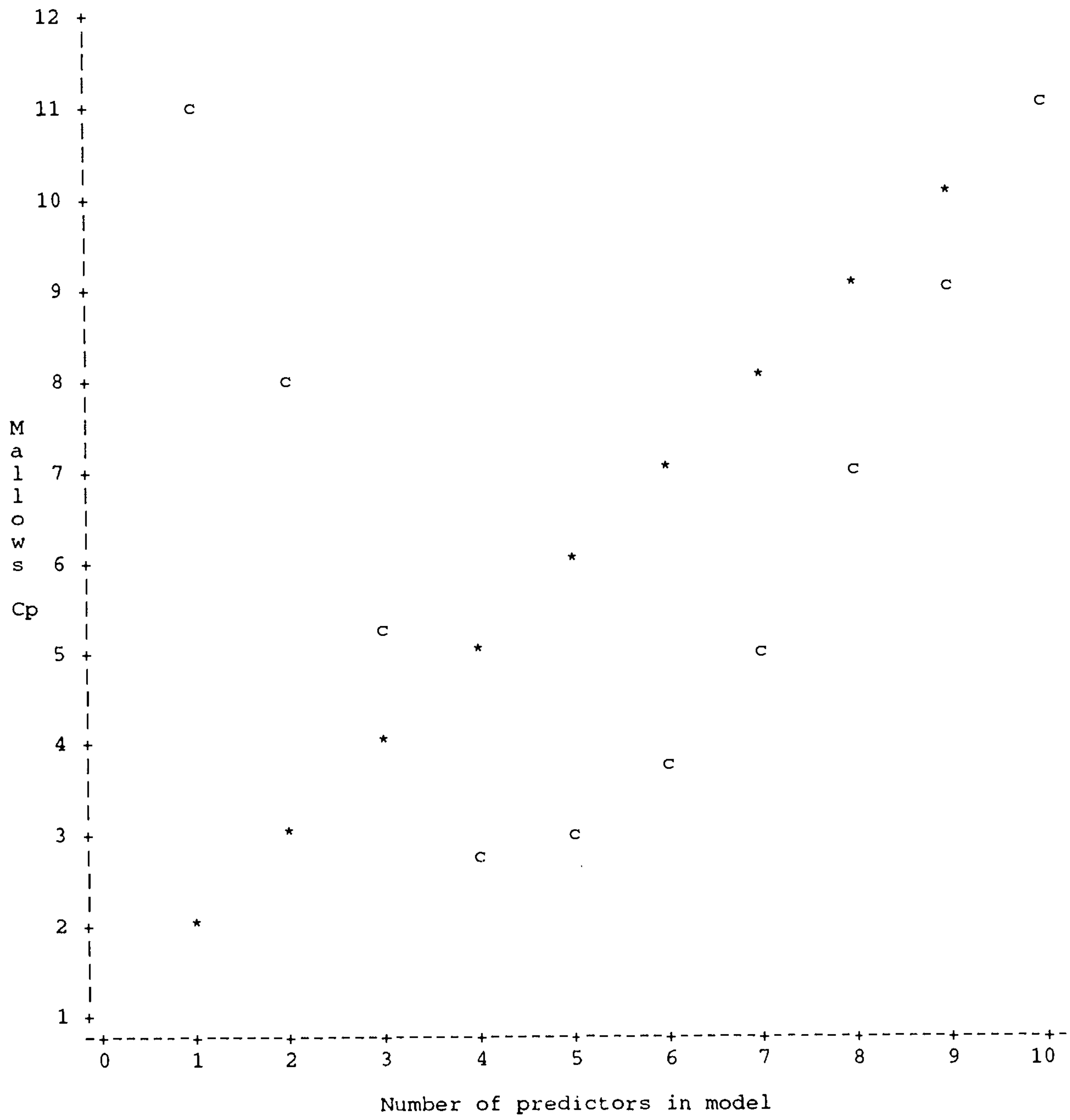
Table 5. Principal Component Estimates

Component	df	Estimate	SD Error	t	p
1	1	.08	.02	3.60	.001
2	1	.06	.04	1.76	.081
3	1	- .08	.04	-1.88	.062
4	1	.14	.06	2.45	.015
5	1	- .03	.06	- .53	.600
6	1	.05	.06	.84	.404
7	1	.10	.08	1.34	.182
8	1	- .18	.09	-2.03	.044
9	1	.09	.09	.95	.341
10	1	- .03	.10	- .31	.758

Table 6. Select Principal Component Eigenvectors

LASSI Subscale	(1)	(4)	(8)
Attention	.31	.47	.13
Motivation	.35	-.23	.13
Time Management	.35	-.04	.26
Anxiety/Worry	.18	.25	-.13
Concentration	.39	-.06	.25
Information	.24	.59	.14
Select Ideas	.30	-.48	-.03
Support	.30	-.28	.40
Class Prep.	.36	.06	-.78
Test Strategy	.33	-.02	.03

Figure 1. Overlay plot of C_p and (p + 1) values.



APPENDIX: SAS STATISTICAL PROGRAM

```
DATA LASSI;INFILE 'A:\LASSI.DAT';IF STATUS=1;
INPUT SEX 27 SATM 31-33 SATV 35-37 STATUS 55 CGPA 67-71
      #2 (Q1-Q77) (77*1.0) #3;
LABEL CGPA = 'FIRST SEMESTER COLLEGE GPA';
*  STATUS   1= 'CURRENT STUDENT' 2= 'WITHDREW';
*  SEX      1= 'FEMALE' 2= 'MALE';
PREATT =  Q5 + Q14 + Q18 + Q29 + Q38 + Q45 + Q51 + Q69;
PREMOT = Q10 + Q13 + Q16 + Q28 + Q33 + Q41 + Q49 + Q56;
PRETMT =  Q3 + Q22 + Q36 + Q42 + Q48 + Q58 + Q66 + Q74;
PREANX =  Q1 +  Q9 + Q25 + Q31 + Q35 + Q54 + Q57 + Q63;
PRECON =  Q6 + Q11 + Q39 + Q43 + Q46 + Q55 + Q61 + Q68;
PREINP = Q12 + Q15 + Q23 + Q32 + Q40 + Q47 + Q67 + Q76;
PRESMI =  Q2 +  Q8 + Q60 + Q72 + Q77;
PRESTA =  Q7 + Q19 + Q24 + Q44 + Q50 + Q53 + Q62 + Q73;
PRESFT =  Q4 + Q17 + Q21 + Q26 + Q30 + Q37 + Q65 + Q70;
PRETST = Q20 + Q27 + Q34 + Q52 + Q59 + Q64 + Q71 + Q75;
LABEL          PREATT = 'ATTITUDE AND INTEREST'
               PREMOT= 'MOTIVATION'
               PRETMT= 'TIME MANAGEMENT'
               PREANX= 'ANXIETY AND WORRY'
               PRECON= 'CONCENTRATION AND ATTENTION'
               PREINP= 'INFORMATION PROCESSING'
               PRESMI= 'SELECT MAIN IDEAS'
               PRESTA= 'SUPPORT TECHNIQUES AND MATERIALS'
               PRESFT= 'SELF TESTING AND CLASS PREPARATION'
               PRETST= 'TEST STRATEGIES';
PROC REG OUTEST=EST; MODEL CGPA = PREATT--PRETST/
          SELECTION = RSQUARE ADJRSQ CP MSE BEST=1;
PROC PLOT;PLOT _CP_ * _IN_ = 'c' _P_ * _IN_ = '*'/OVERLAY
          VAXIS= 0 TO 12 BY 1 HAXIS = 0 TO 10 BY 1;
PROC PRINCOMP DATA=LASSI OUT=PRIN;VAR PREATT--PRETST;
PROC REG; MODEL CGPA = PRIN1-PRIN10/SS2;
RUN;
```