Pak. J. Statist.
2007 Vol. 23 (1) pp 65-81

A MONTE CARLO STUDY OF UNIVARIATE VARIABLE SELECTION CRITERIA

Ali A. Al-Subaihi
College of Education and Humanities
Taibah University
Madinah Al-Monawwarh, Saudi Arabia
Email: alialsubaihi@yahoo.com

ABSTRACT

A Simulation study was conducted, to compare a number of univariate variable selection criteria that are available in either SAS or SPSS, in terms of their ability to select the *"true"* regression model for different sample sizes, intercorrelations, and interacorrelations. The results suggest that the ability of all procedures to identify the *"true"* model is less than 19%, and that sample size, intercorrelations, and interacorrrelations have no significant effect. The study also shows that all criteria are more likely to overfit by at least two variables, than to select the *"true"* model or underfit.

KEYWORDS:

Univariate variable selection, Multiple linear regression, Constructing regression models in SAS and SPSS.

1. INTRODUCTION

Multiple linear regression (MLR), deals with issues related to estimation, or prediction of the expected value of the dependent variable (y), using the known values of (k) predictors (x's). The statistical model of MLR is:

$$y_{n\times 1} = X_{n\times(k+1)}b_{(k+1)\times 1} + \boldsymbol{e}_{n\times 1} \tag{1.1}$$

where $\boldsymbol{y}$ is a vector of responses of n (independent) observations, $\boldsymbol{X}$ is the design matrix of rank k+1, $\boldsymbol{b}$ is a vector of parameters to be estimated or predicted, and $\boldsymbol{e}$ is the vector of residuals.

Theoretically, useful MLR model(s) should be built, based on theories within the field of the problem being studied. However, generally, this is not what is done in practice. What happens is that many researchers use all on hand variables to build the wanted MLR model(s). The justification for this is that the variables are available (Stevens, 1992) and one could apply some of the variable selection methods to identify the *"true"* model.

What practically happens is not necessarily helpful for two reasons: inclusion of too many predictors will cause loss of precision in the estimation of regression coefficients

and the prediction of new responses (Murtaugh, 1998), each variable selection criterion has a different ability to select the *"true"* model from the pool of alternatives under dissimilar situations (McQuarrie and Tsai, 1998; Fujikoshi and Satoh, 1997). Without getting into the philosophical debate about *"true"* model existence in real-life applications, one could define the *"true"* model as the one that contains only the meaningful predictors which best predict future observations.

Despite the fact that statistical literature in variable selection is full of studies that propose either new or modified criteria, there is no empirical study that illustrates the behavior of well-known variable selection criteria, in selecting the *"true"* model under various conditions. Therefore, it is imperative to conduct such a study, especially for new MLR users.

Variable selection criteria available in SAS or SPSS are compared in this paper, in terms of their abilities to select the "*true*" model. These variable selection criteria are particularly chosen, because some of them are usually used in any MLR problem and they cover various criteria, typically discussed in most, if not all, MLR textbooks.

The article is organized as follows. Section 2 lists variable selection criteria that are available in SAS or SPSS and describe some of them briefly. Section 3 reviews some classical and comprehensive work in univariate variable selection. Section 4 provides a description of simulation design, data generation and measures of interest. Section 5 demonstrates the adequacy of data generation and comparison of methods. Section 6 presents conclusions and some final advices.

2. VARIABLE SELECTION CRITERIA

Variable selection criteria can be broadly divided according to their selection mechanic and outcome forms, into two branches: Automatic Search Procedure (ASP) and All-Possible-Regression Procedure (APRP).

ASP develops a sequence of regression models and at each step, it adds or deletes an independent variable (x), based on F statistic and ends with identification of a *single* model as "*true*". The Forward Selection (FS), Backward Elimination (BE), and Stepwise Procedure (SP) are three criteria, which belong to ASP and are available in SAS and SPSS. (SAS/STAT User's Guide, 2004; SPSS Base 10.0 Applications Guide; 1999).

In SAS and SPSS, the FS method calculates the F statistic that reflects the variable's contribution to the model, if it is included, and adds the variable with the largest F statistic that has a significance level greater than 0.5 and 0.05, respectively (SAS/STAT User's Guide, 2004; SPSS Base 10.0 Applications Guide; 1999). SPSS users can also choose to use the F value as a stepping criterion ("F to enter"/ "F to exit"), which does not necessarily provide equivalent results as the significance level does. However, probability significant level criterion is going to be used in this study, because it is the default setting of the stepping criteria for both SAS and SPSS. Furthermore, both packages perform BE by deleting the variables from the model, one by one, until all the variables remaining in the model produce an F statistic significant at the level of 0.1 (SAS/STAT User's Guide, 2004; SPSS Base 10.0 Applications Guide; 1999). Again, SPSS users can also choose to use the F statistic for removing variables from the model. However, this criterion will not

be used. SAS performs SP, following the same idea of the FS and BE procedures, but using different default values (SAS/STAT User's Guide, 2004). SPSS, on the other hand, uses the same idea and default values for entry and removal variables as described in the FS and BE methods (SPSS Base 10.0 Applications Guide; 1999).

In contrast to ASP, the APRP procedure calls for considering all possible subsets of the pool of potential *x's* and identifying a few *good* models according to some criterion. The Coefficient of Multiple Determination (R^2) and Adjusted Coefficient of Multiple Determination (aR^2) are frequently used criteria in APRP and are available in SAS and SPSS (SAS/STAT User's Guide, 2004 and SPSS Base 10.0 Applications Guide, 1999). SAS has nine more APRP criteria: Mallow's C_p, Akaike's Information Criterion (AIC), Sawa's Bayesian Information Criterion (BIC), Estimated Mean Square Error of Prediction (GMSEP), Final Prediction Error (J_p), Mean Square Error (MSE_p), Amemiya's Prediction Criterion (PC), SBC statistic, and S_p index (SAS/STAT User's Guide, 2004).

Both aR^2 and SBC procedures provide equivalent information to MSE_p and BIC procedures, respectively. Therefore, the later two criteria will not be presented here. Moreover, because of space limitations, the mathematical definitions of all the above-listed criteria are introduced briefly. Interested readers are referred to SAS/STAT User's Guide (2004) or Miller (2002) for more details about each criterion.

Before proposing a brief description of each criterion, it would be helpful to introduce a standard notation for all variables, vectors, matrices, and functions used. The following table presents notations and definitions of variables and functions used in defining the criteria:

Table 1
Notations and definitions of variables and functions used.

Symbol	Definition
$\boldsymbol{n}$	The number of observations
$\boldsymbol{p}$	The number of parameters including the intercept
$\boldsymbol{k}$	The number of x's in the "full model"
y	The vector of dependent variable
X	The matrix of all candidate independent variables with its first column being the vector of ones
$\boldsymbol{SSE_{k+1}}$	The sum squared error for a "full-model" including the intercept
$\boldsymbol{SSE_p}$	The sum squared error for a model with p parameters including the intercept
$\boldsymbol{MSE_{k+1}}$	The mean of squares error for a "full-model" including the intercept
$\boldsymbol{ln}(\bullet)$	The natural logarithm

2.1 Coefficient of Multiple Determination $\left(R_p^2\right)$

The R_p^2 index is the proportion of total (corrected) sum of squares accounted for by regression and used as a measure of model fit. For each model, R_p^2 calculated as

$$R_p^2 = 1 - \frac{SSE_p}{SSE_{k+1}}$$

The value of R_p^2 increases as *p* increases and reaches its maximum when all *x's* are in the model. Therefore, we choose a model with a value of *p* beyond which the increases in R_p^2 appear to be unimportant as a *"true"* model. The judgment is subjective and cannot be programmed, therefore, the criterion was omitted from the simulation.

2.2 Adjusted Coefficient of Multiple Determination $\left(aR_p^2\right)$,

aR_p^2 has been suggested as an alternative criterion to R_p^2 and calculated as follows:

$$aR_p^2 = 1 - \frac{(n-1)(1-R_p^2)}{n-p}$$

Like R_p^2, aR_p^2 is calculated for all subsets, but the model that contains meaningful predictors is the one that has the smallest value (Neter et al, 1996).

2.3 Mallow's C_p

The C_p criterion, which was initially introduced by Mallow (1973), is concerned with the total mean squared error of the *n* fitted values for each subset model. It is computed as

$$C_p = \frac{SSE_p}{MSE_{k+1}} - (n - 2p)$$

C_p values below the line $C_p = p$ are interpreted as showing no bias and being below the line due to sampling error. In other words, the *"true"* model is the one that has: (1) small C_p value, and (2) C_p value near *p* (Neter et al, 1996).

2.4 Akaike's Information Criterion (AIC)

The AIC_p procedure (Akaike, 1969) that is often used in practice is considering the Kullback-Leibler (K-L) distance, which is the distance between the true density and estimated density for each model. The K-L measure provides a way to evaluate how well the candidate model approximates the true model by estimating the difference between the expectations of the vector *y* under the true model and the candidate model. The AIC procedure is computed as

$$AIC_p = n \ \ln\left(\frac{SSE_p}{n}\right) + 2p$$

The *"true"* model is the one that is associated with the smallest AIC_p value (the reader is referred to McQuarrie & Tsai (1998) for more statistical details on the AIC_p procedure and its modified forms).

2.5 Estimated Mean Square Error of Prediction (GMSEP)

The $GMSE_p$ technique is

$$GMSE_p = \frac{(n+1)(n-2)}{n(n-p-1)} MSE_P$$

The procedure finds the *"true"* model by computing the estimated mean square error of prediction for each model assuming that both $\boldsymbol{y}$ and $\boldsymbol{x}$*'s* are multivariate normal. Researchers take the model with minimum GMSEP value as the *"true"* model.

2.6 Final Prediction Error (J_p)

The J_p statistic is computed as

$$J_p = \frac{(n+p)}{n} MSE_p$$

where J_p statistic is the estimated mean of squares error of prediction for each model selected assuming that the values of the regressions are fixed and that the model is "true" (SAS/STAT User's Guide, 2004). The statistic is also called the Final Prediction Error (FPE) by Akaike. For more details, the reader is referred to Judge et al. (1980). The "*true*" model is the one that has the minimal value.

2.7 Amemiya's Prediction Criterion (PC)

In order to include a consideration of the losses associated with choosing an incorrect model, Amemiya (1976) developed a criterion based on the square prediction error. The PC_p index is computed for each subset as

$$PC_P = \frac{(n+p)}{(n-p)}\left(1 - R_p^2\right)$$

where R_p^2 is the coefficient of multiple determination. The "*true*" model is the one that corresponds the smallest value.

2.8 SBC statistic

The SBC_p statistic, which is also known as Schwarz's Bayes Information Criterion, was introduced by Schwarz (1978) and uses Bayesian arguments. The statistic is computed as

$$SBC_p = n \ln\left(\frac{SSE_p}{n}\right) + p \ln(n)$$

The SBC index assumes a priori probability of the true model being *p* and a prior conditional distribution of the parameters given that *p* is the true model. Under this assumption, the Bayes solution consists of the model that is a posterior most probable.

That is, the subset that should be included in the model is the one that minimize the SBC_p statistic (for more details, readers are referred to Judge et al. 1980).

2.9 S_p index

The procedure computes the average prediction mean squared error (S_p statistic) for each model selected using mean square error as follows:

$$S_p = \frac{MSE_p}{n - p - 1}.$$

The "*true*" model is the one that has the smallest value.

3. REVIEW OF THE LITERATURE

MLR usually involves a comparison of several candidate models, because the true model is seldom, if ever, known *a priori*. Therefore, a need for objective data-driven methods to employ in the selection of *"true"* models became an increasingly important topic to applied statisticians and continues to receive considerable attention in the recent statistical literature. (Hocking, 1976; Thompson, 1978; McQuarrie & Tsai, 1998; George, 2000; Kadane & Lazar, 2004). A new, or a modified, criterion regularly appears in one of the statistical journals and is added to the univariate variable selection criteria list.

Univariate variable selection literature is very rich. It would not be practical to cover it thoroughly, or even mention all the methods, in one study. Therefore, some classical and comprehensive works in univariate variable selection are summarized only briefly.

Hocking (1976) and Thompson (1978) did two standard studies that are frequently cited. The primary purpose of Hocking's (1976) paper is to provide a review of the concepts and methods associated with variable selection in linear regression models. It reviews the underlying theory and computational techniques of the automatic search procedure, specifically, forward selection and backward elimination and all-possible-regression procedure (explicitly, MSE_p, R_p^2, aR_p^2, C_p, J_p, the average prediction mean squared error, the standardized residual sum of squared, and the prediction sum of squared). Thompson (1978a and 1978b) studied the topic from different angle. The author reviews, evaluates critically and discusses the computational procedures involved in some of the univariate variable selection criteria's execution. The FB, BE, SR, and C_p procedures are amongst these, which the author calls "the most significant methods". The papers say that the procedure(s) should be recommended differently, depending on whether the independent variables must be considered as fixed, or whether it is possible to regard them as random. In the fixed case, for example, Thompson (1978a and 1978b) recommends the C_p procedure.

McQuarrie and Tsai (1998) give a more detailed discussion, with derivations of some of the all-possible-regression procedures commonly used in univariate regression modeling. (C_p, AIC_p, and SBC_p are among these procedures) They evaluate these variable selection criteria in terms of how well the candidate model approximates to the true model given in (1.1), by estimating the difference between the expectations of the vector *y* under the true model and the candidate model using the K-L measure. McQuarrie and

Tsai (1998) conducted Monte Carlo studies to illustrate the behavior of model selection criteria, using two special case models (a model that has strongly identifiable parameters relative to the error and a model that has weakly identifiable parameters) and three different sample sizes, ranging from small to moderate (n = 15, 35, and 100). They concluded that underfitting is less of an issue than overfitting for multivariate regression when the model has strongly identifiable prarmeters, whereas underfitting becomes the main concern when model parameters are only weakly identifiable. Moreover, they stated that some of the classical selection criteria (specifically, C_p and AIC_p) consistently performed poorly, and are not recommended for use in practice.

George (2000) reviews some of the key developments that have led to the wide variety of approaches to the problem of variable selection. The vignette discusses many promising new approaches which have appeared over the last decade, from both frequentist and Bayesian perspectives, in terms of their calculation base and assumptions about which predictor values to use and whether they were fixed or random. The vignette also recommends future works in the field of variable selection. No numerical-based conclusions and recommendations were provided in George (2000).

Miller (2002) describes a very wide range of variable selection techniques in linear regression, which are not necessarily the best. The monograph is intended to alert professional statisticians and advanced students of the field to a class of problems, which arise from the too-routine application of the methods of linear regression, and to the existence of techniques that permit some of those problems to be circumvented. The monograph is an excellent and comprehensive work in variable selection. However, it is a mathematics heavy book with the emphasis on linear algebra which might be uneasy to read for non-statisticians.

The behavior of commonly used univariate variable selection procedures to select the *"true"* model under realistic data conditions, has not been thoroughly documented. Which variable selection criterion selects the *"true"* model more frequently than others, in case of a design consisting of three levels of sample sizes, two levels of correlation among the $\boldsymbol{x}$*'s*, and two levels of correlation between $\boldsymbol{y}$ and the $\boldsymbol{x}$*'s*? This study will compare the ability of the above-listed variable selection criteria to select the *"true"* model under these conditions.

4. METHODOLOGY

4.1 Experimental Design

A Monte Carlo method was used to compare the relative performance of several univariate variable selection procedures in terms of their ability to identify the *"true"* MLR model under realistic conditions. This simulation examines the effects of three manipulated factors under controlling three others, besides the usual controllable statistical MLR assumptions. The controlled factors are the number of the meaningful predictors (m =3), the total number of the *x's* in the full model (k), and the correlation between y and the seGt of useless predictors ($\boldsymbol{S}_2$) (ρ_{yS_2}). The manipulated factors are the sample size (n) (3 levels), the correlation among the $\boldsymbol{x}$*'s* (ρ_x) (2 levels), and the

correlation between y and the set of meaningful predictors ($\boldsymbol{S}_1$) ($\rho_{y\boldsymbol{S}1}$) (2 levels). The simulation study is a $3 \times 2 \times 2$ fully crossed factorial design.

The dependent variable y is assumed to be influenced by two sets of predictors: $\boldsymbol{S_1}$ that contains the *meaningful* predictors and $\boldsymbol{S_2}$ that contains the *useless* noise predictors. The meaningful predictors are highly correlated with y and lowly correlated with each other, and the useless noise predictors that are lowly correlated with y. This would be considered the operational definition of the "*true*" MLR model.

The design was linked to real world univariate regression problems, to select the values of some of the controlled as well as the manipulated factors.

Table 4-1
Summaries of Real World Data

Source	Field	n	K	ρ_x			ρ_{yX}		
				Av.	Max	Min	Av.	Max	Min
Neter et al, (1996, pp. 241.)	Business	21	2	0.78	0.78	0.78	0.89	0.94	0.84
Neter et al, (1996, pp. 25.)	Medicine	20	2	0.93	0.93	0.93	0.97	0.98	0.97
Neter et al, (1996, pp. 252.)	Business	16	2	0	0	0	0.64	0.89	0.39
Neter et al, (1996, pp. 261.)	Medicine	20	3	0.49	0.92	0.08	0.62	0.88	0.14
Neter et al, (1996, pp. 335.)	Medicine	54	4	0.35	0.47	0.25	0.69	0.86	0.56
Neter et al, (1996, pp. 356.)	Business	26	4	0.26	0.50	0.02	0.56	0.72	0.37
Neter et al, (1996, pp. 357.)	Sociology	25	4	0.21	0.47	0.04	0.55	0.83	0.16
Neter et al, (1996, pp.255.)	Education	24	3	0.38	0.78	0.10	0.69	0.90	0.50
Average		26	3	0.43	0.61	0.28	0.7	0.88	0.49

Notes: n is the sample size; k is the total number of predictors; ρ_x is the correlations among the **x**'s; ρ_{yX} is the correlation between $\boldsymbol{y}$ and the set $\boldsymbol{X}$.

Consistent with the above-summarized data, one could set n = 15, 25, and 50, m = 3, and k = 6. Moreover, $\rho_{x_i x_j}$ = 0.2 and 0.5, $i \neq j$, and $\rho_{y\, x_i}$ = 0.5, and 0.8, i = 1, 2, …, k.

4.2 Data Generation and Measures of Interest

The model (1.1), with known regression parameters ($\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 3$, was assumed to underlie the simulated data, and all data was simulated using the SAS RANNOR function within PROC IML (SAS/IML, 2004). The Al-Subaihi (2004) procedures were used to simulate the data.

A total of 1000 samples were generated for each of the 12 conditions that came from combining each of the three numbers of sample sizes with each of the two levels of correlation among the $\boldsymbol{x}$*'s* and with each of the two levels of correlation between y and the helpful predictors.

The percentage of times the *"true"* model is selected by each method and under each condition is the measure of interest and are therefore tabulated.

5. RESULT

5.1 Adequacy of the Data Generation

The adequacy of data generation was judged, by examining the multivariate normality and the desired correlations between y and the sets $\mathbf{S}_1 = \{x_1, x_2, x_3\}$ and $\mathbf{S}_2 = \{x_4, x_5, x_6\}$ separately, and amongst the x's. Results for $n = 15$, $k = 6$, and $m = 3$ for the univariate and multivariate normality test's results and various correlations are tabulated below for the simulated data.

The results in the Table 5-1, suggest that all variables (y, x_1-x_6) are normally distributed according to Kolmogorov's test statistic of univariate normality and Mardia's test statistic of multivariate normality based on multivariate Skewness and Kurtosis, which introduced in Mardia (1974). The Table 5-2 suggests that pair-wise correlations between y and the $\boldsymbol{x}$'s and among the $\boldsymbol{x}$'s are close to the desired values. The average correlation between the $\boldsymbol{x}$'s was calculated using: $\gamma = \dfrac{\lambda - 1}{o - 1}$

where γ is the average correlation, λ is the largest eigenvalue and o is the number of variables.

Table 5-1
Univariate and Multivariate Normality Tests

			n	Test	MS&K	D	P-value
Univariate Normality Test	Dependent Variable	$\boldsymbol{y}$	1500	Kolmogorov	-	.0085	> .150
	Independent Variables	$\boldsymbol{x}_1$	1500	Kolmogorov	-	.0085	> .150
		$\boldsymbol{x}_2$	1500	Kolmogorov	-	.0112	> .150
		$\boldsymbol{x}_3$	1500	Kolmogorov	-	.0120	> .150
		$\boldsymbol{x}_4$	1500	Kolmogorov	-	.0101	> .150
		$\boldsymbol{x}_5$	1500	Kolmogorov	-	.0129	> .150
		$\boldsymbol{x}_6$	1500	Kolmogorov		.0129	> .150
Multivariate Normality Test			1500	Mardia Skewness	.1724	71.954	.8227
			1500	Mardia Kurtosis	63.505	1.1252	.2605

Note: n = The total number of observations obtained from generating 100 samples of size 15; Test = Univariate and multivariate normality Tests; MS&K = Mardia's test statistic of multivariate normality based on multivariate Skewness and Kurtosis; D = Kolmogorov's test statistic of univariate normality; P-value = P-values of the Kolmogorov tests which is < .01 or > .15.

Table 5-2
The Theoretical and Empirical Correlations Values

Theoretical Values			Empirical Values		
ρ_{yx}		ρ_x	ρ_{yx}		γ
ρ_{YS2}	ρ_{yS1}		ρ_{yS2}	ρ_{yS1}	
0.05	0.5	0.2	0.053	0.498	0.199
		0.5	0.044	0.502	0.493
	0.8	0.2	0.043	0.750	0.239
		0.5	0.038	0.799	0.491

Note: ρ_{yx} is the correlation between $\boldsymbol{y}$ and the $\boldsymbol{x}$*'s*; ρ_x is the correlation among the *x's*; ρ_{yS1} is the correlation between $\boldsymbol{y}$ and the set of significant predictors ($\boldsymbol{S}_1$); ρ_{yS2} is the correlation between $\boldsymbol{y}$ and the set of insignificant predictors ($\boldsymbol{S}$); γ is the average correlation among the $\boldsymbol{x}$*'s*.

5.2 Comparison of Methods

In the study, the *"true"* model is the one that includes *only* the predictors x_1, x_2, and x_3, ' underfitted' is one that contains any subset of $\{x_1, x_2, x_3\}$, 'overfitted' is one that includes the set $\{x_1, x_2, x_3\}$, plus any additional predictor(s), and the 'wrong' model is one that contains a *subset* of $\{x_1, x_2, x_3\}$, along with any other predictor(s), or contains any unaccompanied subset of $\{x_4, x_5, x_6\}$.

The results, in Tables 5-3, 5-4, 5-5, and 5-6 suggest that:

1) The ability of each variable selection procedure to select the *"true"* model is quite low; it rangers between 0 to 19% under all conditions. There is no significant difference between SAS default setting of SP, FS, and BE procedures (F= 0.73<3.28 = ${}_{0.95}F_{2,32}$) and SPSS default setting (F= 0.001<3.28 = ${}_{0.95}F_{2,32}$), in terms of their ability to select the *"true"* model. Furthermore, there is also no significant difference between the group of ASP procedures in SAS and in SPSS (T^2 = 7.13<10.22 = ${}_{0.05}T^2{}_{3,22}$, where T^2 is *Hotelling's* T^2 test). On the other hand, all-possible-regression-procedures (APRP) are not significantly better than ASP. That is, there is no significant difference between the group of APRP and the group of ASP (T^2 = 8.021<20.95 = ${}_{0.05}T^2{}_{6,22}$), and within APRP (F= 0.14<2.12 = ${}_{0.95}F_{7,88}$) in terms of their selection ability of the right model. [The estimated mean square error of the prediction procedure (GMSE), is equivalent to the average prediction mean squared error method (Sp), and the Amemiya's prediction criterion procedure (Pc), is equivalent to the final prediction error method (Jp). Thus, Sp and Jp methods were temporarily deleted to perform *Hotelling's* T^2 test]. The sample size, correlation among the **x**'s, and correlation between the **x**'s and **y** do *not* have a significant role in increasing (or decreasing) the criterion's ability to select the right model.

2) All variable selection procedures (with no exceptions), are more likely to underfit by at least one variable than to select the *"true"* model. In some conditions, the probability of some selection methods to underfit is approximately 80%. Under all circumstances, the possibility of ASP in SPSS to underfit is significantly higher than the possibility of ASP in SAS (T^2 = 68.14>10.22 = $_{0.05}$ $T^2_{3,22}$). In detail, forward selection method in SAS and backward elimination in SPSS, are less likely to underfit than other ASP criteria do. On the other hand, APRP are significantly less likely to underfit the model than ASP do (T^2 = 90.72>20.95 = $_{0.05}$ $T^2_{6,22}$). The sample size, correlation among the **x**'s, and correlation between the **x**'s and **y** *do* play a vital role in decreasing the probability of underfitting. That is, as n, ρ_x, or ρ_{xy} increases, the probability of underfitting decreases.

3) Once again, every variable selection procedure is more likely to overfit by at most two variables than to select the *"true"* model. In detail, under all circumstances, the possibility of ASP in SAS to overfit is significantly higher than the possibility of ASP in SPSS (T^2 = 44.18>10.22 = $_{0.05}$ $T^2_{3,22}$). More exactly, stepwise procedure, (SP) in SAS and SPSS are less likely to overfit than FS and BE methods do. Conversely, APRP have significantly more chance of overfit than the ASP (T^2 = 177.51>20.95 = $_{0.05}$ $T^2_{6,22}$). Furthermore, n, ρ_x, and ρ_{xy} have a positive effect on the probability of overfitting (i.e., large values of n, ρ_x, and ρ_{xy} increase the probability of overfitting).

Surprisingly, most of the time, the probability of all variable selection criteria to select the wrong model is quite high. The average probability of selecting the *"true"* model wrongly, is around 47%, nevertheless, in some conditions the possibility reaches 84 %. The probability that automatic search procedures in SAS will select the wrong model is significantly higher than the possibility of ASP in SPSS in most circumstances (T^2 = 17.51>10.22 = $_{0.05}$ $T^2_{3,22}$). In some conditions, For example, at of n = 50, ρ_x = 0.2 and ρ_{xy} = 0.5 and 0.8, ASP methods in SAS are less likely to select the wrong model than in SPSS. Furthermore, APRP procedures are significantly more likely to select the wrong model than ASP procedures (T^2 = 133.70>20.95 = $_{0.05}$ $T^2_{6,22}$).

Table 5.3
The Percentage (%) of Times the *"true"* Model is Selected

n	ρ_x	ρ_{xy}	Selection Criteria														
			RAN	ASP in SAS			ASP in SPSS			APRP							
				SP	FS	BE	SP	FS	BE	AR	AIC	GMSE	Jp	Pc	SBC	Sp	Cp
15	.2	.5	1.90	2.10	0.80	1.10	0.20	0.80	0.20	1.80	1.70	1.10	1.90	1.90	0.90	1.10	2.34
		.8	1.50	4.30	2.80	5.60	2.30	2.80	1.80	5.10	5.50	5.80	5.90	5.90	5.50	5.80	3.49
	.5	.5	1.30	0.80	0.20	0.30	0.00	0.20	0.00	0.40	0.30	0.30	0.30	0.30	0.20	0.30	1.41
		.8	1.60	0.40	0.20	0.20	0.00	0.20	0.00	0.30	0.30	0.30	0.30	0.30	0.30	0.30	1.02
25	.2	.5	1.40	3.80	1.50	2.80	1.10	1.50	1.10	4.90	4.50	3.60	4.50	4.50	1.50	3.60	3.35
		.8	1.40	2.90	6.80	8.00	5.80	6.80	5.40	5.60	7.80	7.70	8.20	8.20	6.80	7.70	4.58
	.5	.5	1.40	0.10	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	1.20
		.8	2.10	0.10	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.23
50	.2	.5	1.70	3.00	10.00	11.70	8.10	10.00	8.10	7.90	10.90	11.20	10.90	10.90	8.80	11.20	5.50
		.8	1.20	1.20	13.90	9.60	18.50	13.90	18.50	3.00	7.60	8.40	7.60	7.60	16.20	8.40	4.26
	.5	.5	2.20	0.00	0.10	0.20	0.00	0.10	0.00	0.00	0.10	0.10	0.10	0.10	0.00	0.10	0.32
		.8	1.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: n is the sample size; ρ_x is the correlation among the *x*'s; ρ_{xy} is the correlation between the *x*'s and y; RAN stands for random selection (i.e., applying no selection criteria); ASP is automatic search procedures; SP is the stepwise procedure; FS is the forward selection method; BE is the backward elimination; APRP is the all-possible-regression procedures; AR is the adjusted R-square; AIC is the Akaike's information criterion; GMSE is the estimated mean square error of prediction; Jp is the final prediction error; Pc is the Amemiya's prediction criterion; SBC is the Schwarz's Bayes information criterion; Sp is the average prediction mean squared error; Cp is the Mallow's Cp criterion.

Table 5.4
The Percentage (%) of Times the Criteria Underfit by at Least One Variable

n	ρ_x	ρ_{xy}	Selection Criteria														
			RAN	ASP in SAS			ASP in SPSS			APRP							
				SP	FS	BE	SP	FS	BE	AR	AIC	GMSE	Jp	Pc	SBC	Sp	Cp
15	.2	.5	8.70	5.90	47.40	51.20	80.20	47.40	80.20	16.60	28.90	40.30	30.80	30.80	41.70	40.30	8.89
		.8	9.40	3.70	43.60	40.10	72.70	43.60	73.20	11.40	19.90	30.70	21.70	21.70	30.70	30.70	7.33
	.5	.5	8.00	1.90	25.90	31.90	61.30	25.90	61.30	5.30	13.30	20.30	14.30	14.30	21.80	20.30	6.74
		.8	9.20	0.30	11.80	13.70	42.50	11.80	42.50	2.10	3.70	7.50	4.20	4.20	8.00	7.50	3.20
25	.2	.5	10.20	3.70	48.20	44.40	76.20	48.20	76.30	12.00	28.00	36.30	28.70	28.70	52.40	36.30	7.68
		.8	9.50	0.60	30.90	23.80	56.00	30.90	56.40	3.50	12.20	18.80	12.90	12.90	31.00	18.80	4.31
	.5	.5	9.10	0.90	18.60	18.70	50.20	18.60	50.20	3.00	9.50	11.00	9.60	9.60	23.20	11.00	5.17
		.8	10.20	0.10	1.70	1.10	13.00	1.70	13.00	0.10	0.30	0.60	0.50	0.50	1.90	0.60	0.46
50	.2	.5	9.90	1.50	29.50	20.90	58.10	29.50	58.10	3.90	15.20	17.80	15.40	15.40	49.20	17.80	4.94
		.8	9.80	0.00	4.00	2.10	14.60	4.00	14.60	0.00	1.20	1.70	1.20	1.20	10.10	1.70	0.86
	.5	.5	9.00	0.10	3.00	3.20	17.30	3.00	17.30	0.20	1.00	1.40	1.10	1.10	8.50	1.40	0.92
		.8	9.80	0.00	0.00	0.00	0.30	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: n is the sample size; ρ_x is the correlation among the *x*'s; ρ_{xy} is the correlation between the *x*'s and y; RAN stands for random selection (i.e., applying no selection criteria); ASP is automatic search procedures; SP is the stepwise procedure; FS is the forward selection method; BE is the backward elimination; APRP is the all-possible-regression procedures; AR is the adjusted R-square; AIC is the Akaike's information criterion; GMSE is the estimated mean square error of prediction; Jp is the final prediction error; Pc is the Amemiya's prediction criterion; SBC is the Schwarz's Bayes information criterion; Sp is the average prediction mean squared error; Cp is the Mallow's Cp criterion.

Table 5.5
The Percentage (%) of Times the Criteria Overfit by at Most Two Variables

n	ρ_x	ρ_{xy}	Selection Criteria														
			RAN	ASP in SAS			ASP in SPSS			APRP							
				SP	FS	BE	SP	FS	BE	AR	AIC	GMSE	Jp	Pc	SBC	Sp	Cp
15	.2	.5	10.50	25.00	3.10	2.20	0.40	3.10	0.40	14.10	9.50	4.50	7.90	7.90	5.60	4.50	17.16
		.8	8.60	38.30	5.10	7.80	1.50	5.10	1.00	27.30	19.90	10.20	17.60	17.60	12.60	10.20	28.83
	.5	.5	10.50	21.40	1.80	1.50	0.30	1.80	0.00	13.30	8.80	3.50	7.10	7.10	4.60	3.50	16.97
		.8	8.00	31.90	4.00	4.70	1.20	4.00	0.50	23.00	16.60	7.50	14.20	14.20	9.90	7.50	26.00
25	.2	.5	9.00	39.10	4.60	6.50	1.00	4.60	0.80	25.90	13.00	8.90	11.70	11.70	4.70	8.90	27.69
		.8	8.60	60.10	19.60	25.60	7.10	19.60	6.30	54.40	38.40	30.30	36.80	36.80	19.70	30.30	47.84
	.5	.5	11.40	32.60	5.50	6.20	0.90	5.50	0.80	22.30	12.60	8.80	12.00	12.00	6.20	8.80	23.32
		.8	8.30	36.90	17.60	21.00	7.90	17.60	5.90	41.90	34.00	25.70	32.80	32.80	17.60	25.70	40.13
50	.2	.5	9.00	61.30	27.10	34.10	9.30	27.10	9.00	60.90	42.60	39.00	42.30	42.30	13.40	39.00	51.23
		.8	8.70	57.80	69.30	73.90	46.40	69.30	46.20	74.50	76.50	76.50	76.60	76.60	54.80	76.50	77.12
	.5	.5	9.00	36.50	26.70	31.10	11.10	26.70	10.20	42.00	37.30	35.70	37.60	37.60	15.60	35.70	40.02
		.8	9.10	20.20	66.50	47.90	41.70	66.50	41.60	30.70	43.70	47.90	44.10	44.10	47.00	47.90	33.28

Note: n is the sample size; ρ_x is the correlation among the *x*'s; ρ_{xy} is the correlation between the *x*'s and y; RAN stands for random selection (i.e., applying no selection criteria); ASP is automatic search procedures; SP is the stepwise procedure; FS is the forward selection method; BE is the backward elimination; APRP is the all-possible-regression procedures; AR is the adjusted R-square; AIC is the Akaike's information criterion; GMSE is the estimated mean square error of prediction; Jp is the final prediction error; Pc is the Amemiya's prediction criterion; SBC is the Schwarz's Bayes information criterion; Sp is the average prediction mean squared error; Cp is the Mallow's Cp criterion.

Table 5.6
The Percentage (%) of Times the Criteria Selected the Wrong Model as a "true" Model

n	ρ_x	ρ_{xy}	Selection Criteria														
			RAN	ASP in SAS			ASP in SPSS			APRP							
				SP	FS	BE	SP	FS	BE	AR	AIC	GMSE	Jp	Pc	SBC	Sp	Cp
15	.2	.5	77.20	59.70	42.40	45.50	19.20	42.40	19.20	65.60	58.40	53.80	58.40	58.40	51.00	53.80	71.52
		.8	79.10	39.90	48.50	46.10	23.40	48.50	23.90	52.50	52.20	52.40	53.40	53.40	50.10	52.40	60.02
	.5	.5	78.60	65.40	67.80	66.20	38.40	67.80	38.70	78.00	75.30	75.50	76.80	76.80	72.30	75.50	74.69
		.8	79.80	52.00	83.90	81.20	56.30	83.90	57.00	71.40	77.10	84.10	79.70	79.70	80.40	84.10	69.42
25	.2	.5	77.80	42.50	45.20	46.30	21.70	45.20	21.80	55.50	53.70	51.10	54.60	54.60	41.40	51.10	61.14
		.8	78.70	18.40	42.70	42.30	31.10	42.70	31.90	30.90	40.10	42.70	41.10	41.10	42.10	42.70	42.31
	.5	.5	76.80	50.10	75.10	75.00	48.90	75.10	49.00	69.10	75.40	79.80	76.60	76.60	70.50	79.80	69.73
		.8	78.10	25.10	80.70	75.10	78.90	80.70	81.00	42.20	57.90	69.30	60.00	60.00	77.70	69.30	54.02
50	.2	.5	77.30	12.90	33.40	32.60	24.50	33.40	24.80	20.60	29.60	31.00	29.90	29.90	28.60	31.00	37.11
		.8	78.40	1.60	12.80	9.90	20.30	12.80	20.50	3.50	7.40	8.30	7.50	7.50	18.60	8.30	9.59
	.5	.5	78.90	15.50	70.20	58.50	71.00	70.20	71.90	28.30	49.90	54.40	49.90	49.90	74.80	54.40	46.96
		.8	78.00	2.20	33.50	19.40	46.30	33.50	47.60	6.30	13.10	16.20	13.40	13.40	38.70	16.20	11.85

Note: n is the sample size; ρ_x is the correlation among the $\boldsymbol{x}$'s; ρ_{xy} is the correlation between the $\boldsymbol{x}$'s and y; RAN stands for random selection (i.e., applying no selection criteria); ASP is automatic search procedures; SP is the stepwise procedure; FS is the forward selection method; BE is the backward elimination; APRP is the all-possible-regression procedures; AR is the adjusted R-square; AIC is the Akaike's information criterion; GMSE is the estimated mean square error of prediction; Jp is the final prediction error; Pc is the Amemiya's prediction criterion; SBC is the Schwarz's Bayes information criterion; Sp is the average prediction mean squared error; Cp is the Mallow's Cp criterion.

6. CONCLUSION

This study investigates the ability of a variety of univariate variable selection criteria to select the *"true"* model. It compares the default setting of the automatic search procedures (the stepwise procedure, forward selection, and backward elimination) in SAS and SPSS as well as the commonly used all-possible-regression-procedures (the adjusted R-square, Akaike's information criterion, estimated mean square error of prediction, final prediction error, Amemiya's prediction criterion, Schwarz's Bayes information criterion, average prediction mean squared error, and Mallow's Cp criterion.).

The results of the study suggest that the SAS and SPSS default setting of automatic search procedures and the group of all-possible-regression-procedures have quite a low chance (less than 19%) of selecting the *"true"* model. Moreover, there are a number of procedures that are useless in terms of *"true"* model selection under some conditions. For instance, the ability of all techniques at n = 50, $\rho_x = 0.5$, and $\rho_{xy} = 0.8$ to select the right model is inferior to random selection criterion (RAN), which represents not applying a variable selection technique.

The study shows that probability of a backward elimination procedure in both SAS and SPSS to select the "*true*", the wrong, overfit, or underfit are exactly the same. Similarly, the output of the estimated mean square error of prediction procedure (GMSE) is equal to the output of the average prediction mean squared error method (Sp) and the output of Amemiya's prediction criterion procedure (Pc), which is equal to the output of the final prediction error method (Jp).

The simulation shows that the sample size, amount of correlations among the independent variables, and magnitude of correlations between the dependent and independent variables do *not* influence significantly the ability of any univariate variable selection criterion to select the *"true"* model.

The results show how dangerous it is when the investigator depends on variable selection criteria to select the *"true"* model and pays modest attention to his/her knowledge of the substantive area under study. The study confirms that it is important for the investigator to be judicious in his/her selection of predictors and (1) include predictors based on theory(ies) of the field under study and have low correlation with each other or high correlation with the dependent variable, (2) utilizing more than one criterion in evaluating a possible subset of independent variables, and (3) evaluating the final *"true"* models using various diagnostic procedures before the final regression models are determined.

7. ACKNOWLEDGEMENT

The author is thankful to the reviewers and the Chief Editor for useful comments. These comments were really useful to upgrade the quality of the paper.

REFERENCES

1. Al-Subaihi, Ali A. (2004). Simulating Correlated Multivariate Pseudorandom Numbers. *J. Statist. Software*, 9(4), http://www.jstatsoft.org/index.php?vol=9

2. Fujikoshi, Y., and Satoh, K. (1997). Modified AIC and Cp in Multivariate Linear Regression. *Biometrika*, 84(3), 707-716.
3. George, E.I. (2000). The Variable Selection Problem. *J. Amer. Statist. Assoc.*, 95, 1304-1308.
4. Hocking, R.R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32, 1-49.
5. Kadane, J B., and Lazar, N.A. (2004). Methods and Criteria for Model Selection. *J. Amer. Statist. Assoc.*, 99, 279-290.
6. Mardia, K.V. (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. *Sankhya B*, 36, 115-128.
7. McQuarrie A.D., and Tsai, C. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing Co. Pte. Ltd., River Edge, NJ.
8. Miller, A.J. (2002). *Subset Selection in Regression*, Chapman and Hall, New York, NY.
9. Murtaugh, P.A. (1998). Methods of Variable Selection in Regression Modeling. *Commun. Statist.- Simula*, 27(3), 711-734.
10. Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996). *Applied Linear Statistical Models*. McGraw-Hill Companies, Inc., NY.
11. *SAS/IML Software*, SAS Institute Inc. (2004). *SAS OnlineDoc 9.1.2***.** Cary, NC: SAS Institute Inc.
12. *SAS/STAT User's Guide*, SAS Institute Inc. (2004). *SAS OnlineDoc 9.1.2***.** Cary, NC: SAS Institute Inc.
13. *SPSS Base 10.0 Applications Guide*, SPSS Inc., Chicago, IL (1999).
14. Stevens, J. (1992). *Applied Multivariate Statistics for the Social Sciences* (2nd Edition). Lawrence Erlbaum Association, Publishers, Hillsdale, New Jersey.
15. Thompson, M.L. (1978a). Selection of Variable in Multiple Regression: Part I. A review and Evaluation. *International Statistical Review*, 46, 1-19.
16. Thompson, M.L. (1978b). Selection of Variable in Multiple Regression: Part II. Chosen Procedures Computations, and Examples. *International Statistical Review*, 46, 129-146.