IASC 2008: December 5-8, Yokohama, Japan

A Randomization Approach Utilizing the Ordered Effect Size to Assess Statistical Significance in Forward Selection Models

Yasser Shehata[1] Paul White[2]

[1]Productivity and Quality Institute, Arab Academy for Science and Technology, Egypt.
E-mail: yasser.shehata@live.uwe.ac.uk

[2]Bristol Institute of Technology, The University of the West of England, Bristol BS16 1QY, UK.
E-mail: Paul.White@uwe.ac.uk

Keywords: Forward selection, Randomization, Type I error, Ordered effect size

Abstract

Automated subset selection techniques such as forward selection are commonly used in model building. It is well known that these techniques are problematic. Commonly cited problems include the over capitalization on chance effects leading to problems of the incorrect identification of the true model. Biases in the distribution of p-values and errors in inference arising from standard automated selection procedures are non-trivial and the extent of the problem is dependent on the number of potential predictors and on their correlation structure. The problems arise for using data steered algorithms to uncover potentially good models and to assess the identified models using the same data as if they were prior specified models without regards to the data steering aspect of the algorithm.

We propose a novel randomization algorithm for assessing the statistical inference aspects of forward selection. This proposed algorithm explicitly takes into account the data steering aspect of the algorithm. The proposed algorithm transfers inference from named predictor variables to variables ordered on size of their effect in a discovered model. Unlike traditional forward selection this approach allows the models to be pre-specified. Under a global null model the proposed algorithm is shown to correct the non-trivial biases associated with traditional forward selection and do so irrespective of the number of predictors and irrespective of the correlation structure between predictors. In the non-null situation the proposed algorithm retains power for authentic predictors and minimizes the bias when compared with the traditional approach. The extent of the bias is dependent on the true state of nature however as the number of noise variables increase the biases with the traditional approach increases but the bias in the distribution of the p-values asymptotically decreases under the proposed algorithm. A joint decision rule to curb excessive Type I errors in forward selection is also considered.

1 Introduction

In exploratory analyses using multiple regression it is usual for a researcher to postulate a theoretical model of the form $Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \varepsilon$ and on the basis of sample data to make inferences as to which group of variables form a best subset for the prediction of Y. Analysis usually proceeds on the presumption that the model specification encompasses a true specification and that some variables may have regression weights $\beta_p = 0$ in the true specification. In a non-null situation, the variables that form the true specification and which have non-zero regression weights form the set of authentic variables and all variables outside of this set are non-authentic variables and have regression weights $\beta_p = 0$ in the authentic model (although by themselves they may not be noise variables). The identification of variables to retain in a final model is seen to be equivalent to the identification of the set of authentic variables. In this sense the ideas of variable selection and subset selection become synonymous.

Subset selection in ordinary least squares multiple regression is long established. Computational algorithms for forward selection techniques date back to at least the 1950's (see for instance Kramer, 1957). The use of automated computer subset selection techniques is widespread. Over 25 years ago, Miller (1984) estimated that "approximately 10^5 multiple regressions were carried out each day", with a large number of these using subset selection. More recently, George (2000) writes "The problem of variable selection is one of the most pervasive model selection problems in statistical application. The use of variable selection procedures will only increase as the information revolution brings us larger data sets with more and more variables. The demand for variable selection will be strong and it will continue to be a basic strategy for data analysis".

The motivation for subset selection is well established. Benefits afforded are a parsimonious explanation and greater insight into the phenomenon under investigation with potentially improved predictive accuracy by not including seemingly irrelevant variables in a fitted model. A decision of classifying a noise variable and/or non-authentic variable as an authentic variable may lead to "overfitting" whereby idiosyncratic characteristics in a sample are reflected in a sample based model but which are not replicated on new samples. The non-inclusion of seemingly irrelevant variables in a model partly prevents "overfitting". However it is known that "overfitting" is not eliminated by the use of automated variable selection techniques.

Simulation work of Derksen and Keselman (1992) established that application of standard stepwise techniques may result in 20% to 74% of potential predictors which are scientifically unrelated to the response being classed as statistically significant predictors in a final model. Their work concluded that the use of automated subset selection techniques provided no guarantee that genuine authentic variables would end up in a final model and that the frequency of authentic variables being included in a final model depended on the correlation structure between potential predictors. They also concluded that the frequency with which noise variable were included in the model increased with the number of candidate predictors and that sample size was "of little practical importance in determining the number of authentic variables contained in the final model". However Flack and Chang (1987) found that the use of automated subset selection techniques is "particularly problematic when the sample size is not large relative to the number of candidate variables". Accordingly the problems associated with subset selection are situation dependent.

A major problem with the automated techniques is that classical inference incorrectly proceeds ignoring the data steering in selecting the subset. Classical inference proceeds on the basis that the post hoc identification of the subset is as though it is a prior determined subset leading to an underestimation of the variance associated with the estimated regression weights. This, coupled without explicit control for the multiplicity of tests performed, may lead to an increased number of Type I errors and there is no simple quantification of the probability of the occurrence of Type I errors (see for instance, Derksen and Keselman (1992)). In recognition of this Buckland et al. (1997) have argued that "model selection uncertainty should be fully incorporated into statistical inference".

In the following we consider the control of Type I error rates using a randomization technique with forward selection. Forward selection is investigated not because of any claim of superiority to other techniques but simply because it is very widely used and can also be used in those situations whereby the number of cases exceeds the number of potential predictors (unlike backward elimination). In addition Berk (1978) compared the forward selection algorithm with the best subset from all subsets algorithm using simulation and found that the all subsets approach often identified the same model as found using forward selection but where disagreement occurred the forward selection procedure was superior in identifying the model corresponding to the true state of nature (the all subsets approach can have a tendency to have increased chances of "overfitting").

The paper is organized as follows: In Section 2, we briefly outline the traditional forward selection method and discuss issues in inference using the method. In Section 3, we describe a randomization approach that empirically estimates individual significance of variables selected by forward selection at a given step based on their contribution to the model. In Section 4, a description of two models subsequently used to compare the performance of the randomization algorithm with the traditional forward selection approach is given. Results of the simulation study, the effects of the number of predictors and the effects of sample size are given in Sections 5 and 6. Interpretation of the variable significance under the randomization algorithm can lead to some paradoxical findings. In Section 7 we give a decision rule that combines the results from the traditional approach and the randomization approach which reduces paradoxical findings in practice.

2 Forward Selection

In the forward selection algorithm, the candidate variable for possible inclusion into a forward selection model is that variable not currently in the model which maximizes the squared partial correlation with the dependent variable given the other variables already selected. This candidate variable is then included in the model provided that the partial correlation between the dependent variable and the variable to be added, after allowing for the other variables previously selected, can be considered to be not equal to zero. The algorithm continues until all variables are in the model unless a stopping criterion to conclude a final model is used. In commonly used computer packages this inference is made ignoring the decision making previously made and proceeds on the basis that the current model was not data steered. At stage k of the algorithm, the hypothesis tests are all of the form $H_0\colon \beta_p = 0, (p = 1,2,\dots,P)$ conditional on the other variables being in the model and the p-values are not adjusted in any way to reflect how the model was derived. Thus, the hypothesis tests conducted at a given step will always be tests on random variables. In recognition of this, Pinsker et al. (1987) have considered null hypotheses stating that the coefficients of the remaining regressors not currently included in the model are equal to zero. However it is questionable whether this "lack-of-fit" approach directly answers the question of interest.

Suppose that $X_{[l,k]}$ is the candidate variable entered in to the model at step l of forward selection and its significance is to be judged at stage k. Without loss of generality suppose variable X_1 is $X_{[l,k]}$ $(1 \le l \le k)$. How should the statistical significance of X_1 be judged at stage k? The standard approach would be to statistically test whether the partial correlation coefficient between X_1 and Y given the other variables is equal to zero irrespective of the earlier decision making at previous steps. Since X_1 was not selected in advance then arguably the inference should proceed on $X_{[l,k]}$ and not on X_1. Performing a statistical assessment on $X_{[l,k]}$ would take into account the data steering approach under the algorithm. More generally any specified predictor X_p may not be in a forward selection solution at a particular stage but $X_{[l,k]}$ will, by definition, be in the solution $(l \le k)$. Viewed this way, inference would be shifted from a named variable onto a variable identified by order of entry. In a similar way at step k of the forward selection procedure we may consider a model of the form $Y = \gamma_0 + \gamma_1 V_{[1,k]} + \gamma_2 V_{[2,k]} + \cdots + \gamma_k V_{[k,k]}$, where $V_{[1,k]}$ denotes the variable in model k that has the largest partial correlation with Y, where $V_{[2,k]}$ denotes the variable in model k that has the second largest squared partial correlation with Y and so on. Viewed this way, inference would be shifted from named variables onto variables identified by the relative contribution to a model at step k.

An assessment of the statistical significance of $V_{[l,k]}$ can be made using randomization preserving the problem dependent correlation structure observed between sample predictors. These features of randomization are expanded on in the following sections.

3 Randomization Method

Consider sample data $y_i, x_{1i}, x_{2i}, \ldots, x_{Pi}$, $(i = 1,2,\ldots,I)$ and let $F(l,k)$ denote the F-statistic at step k for variable $V_{[l,k]}$ $(l = 1,\ldots,k; k = 1,\ldots,P)$ so that $F(1,k) > F(2,k) > \cdots > F(k,k)$.

Now consider where the order of cases for the predictor variables in the data is randomly permuted but with the response held fixed i.e. $y_i, x_{1i}, x_{2i}, \ldots, x_{Pi} \to y_i, x_{1m}, x_{2m}, \ldots, x_{Pm}$. This random permutation of predictor records ensures that the sample correlation structure between the predictors in the original data set is precisely preserved in the newly created randomized data set. The random permutation also ensures that the predictor variables in the randomized data set are stochastically independent of the response, Y, but may be correlated with Y in any sample through a chance arrangement. This procedure is logically equivalent to permuting the Y values and holding the predictor variables fixed.

Forward selection can be performed on the newly created randomized data set. Let $F_j(l,k)$ denote the F-statistic at step k for the l-th ordered variable $V_{[l,k]}$ for the j-th randomized data set. The relative ordering of $F(l,k)$ in the empirically determined distribution of $F_j(l,k)$ provides an estimate of the p-value for $V_{[l,k]}$ at step k ($l = 1, \ldots k$; $k = 1, \ldots, P$). The estimate of the p-value for $V_{[l,k]}$ at step k is the estimated proportion of times $F(l,k)$ is lower than $F_j(l,k)$.

4 Design of the Simulation Study

The simulation study was designed so that it reflects selection issues under a global null-model and non null situation as encountered in practice. For specific applications consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

To demonstrate the properties of the proposed technique, two specific parameter settings (referred to in the following as Model A and Model B) with two different situations describing the correlation structures between potential predictors have been considered. Situation (1); stochastically independent predictors in which their correlation matrix is the identity matrix, and Situation (2); strongly correlated predictors with elements of the correlation matrix being $\rho(X_1,X_2) = 0.708$, $\rho(X_1,X_3) = 0.802$, $\rho(X_1,X_4) = -0.655$, $\rho(X_2,X_3) = 0.757$, $\rho(X_2,X_4) = -0.582$, $\rho(X_3,X_4) = -0.593$, where $\rho(X_l,X_m)$ denotes Pearson's correlation coefficient.

In the following simulations, Model A is a genuine null model with $\beta_0 = 1$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, i.e. all proposed predictors are in fact noise variables both in situation (1) and (2). For Model B a genuine non-null model we consider, $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = \beta_3 = \beta_4 = 0$, i.e. one authentic variable under situation (1) and (2), and three noise variables under situation (1) and three non-authentic variables under situation (2). In all instances the error terms are independent identically distributed realizations from the standard normal distribution $(\mu = 0, \sigma^2 = 1)$ so that the underpinning assumptions for the OLS linear regression models are satisfied.

For both Models A and B we have considered $n = 1{,}000$ realization of sample size $I = 30$ from each of Situation 1 (uncorrelated predictors) and Situation 2 (correlated predictors). For each data set generated we have created $J = 1{,}000$ randomized data sets to estimate the p-values of $V_{[l,k]}$ $(l = 1, \dots, k; k = 1, \dots, 4)$. In what follows simulations are reported based on $I = 30$ cases per simulation instance and we later consider increasing sample size and increasing the number of potential predictor.

5 Results of the Simulation Study

In the simulation analysis we initially considered $n = 1{,}000$ data sets each of size $I = 30$ drawn from the aforementioned theoretical model under situation 1 (uncorrelated predictors). In each instance the traditional forward selection algorithm was performed using $k = 4$ steps. At each stage of each forward selection run the standard p-value for each variable in each model was recorded. Thus for instance at step 3, three conditional p-values were obtained and indexed $V1|V2, V3$, $V2|V1, V3$, and $V3|V1, V2$ for $V_{[1,3]}$, $V_{[2,3]}$, $V_{[3,3]}$ respectively. Figure 1 is a percentile plot of the observed p-values for $V1|V2, V3$, $V2|V1, V3$, and $V3|V1, V2$ against the expected percentiles of the uniform distribution $U(0,1)$, at stage $k = 3$. Note that the distribution of the p-values in this arrangement are not uniformly distributed, so that if a researcher worked against an assumed nominal significance level α then the true size of the test would not be α. The distribution of the p-values for $V1|V2, V3$ and $V2|V1, V3$ is clearly stochastically smaller than $U(0,1)$ whereas the distribution of the p-values for $V3|V1, V2$ is stochastically larger than $U(0,1)$. Similar non-uniformity was observed at other steps of the forward selection procedure.

For each data set, randomization was performed $(J = 1{,}000)$ and the associated p-values for $V_{[l,k]}(l = 1, \dots, 3; k = 1, \dots, 4)$ was calculated. In all instances the empirical distribution of the p-values were consistent with the uniform distribution $U(0,1)$, (see Figure 1). Repeating the process under Model A, but with situation 2 (highly correlated predictors), gave the same conclusions for the randomization algorithm and broadly the same conclusions for the traditional approach.

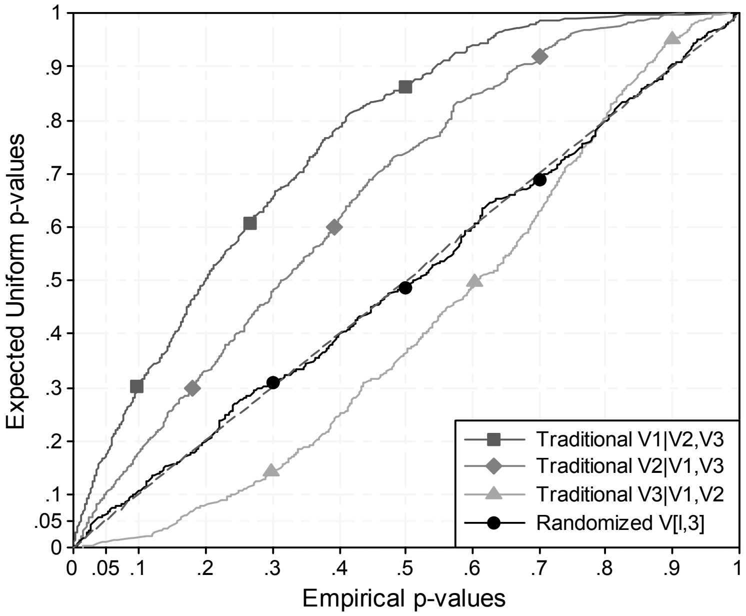

Figure 1: Percentile-Percentile Plot for p-values at Step $k = 3$ for Traditional Approach and for Randomized approach for $V_{[l,3]}$ under Model A.

Simulations under Model B for step $k = 2$ in forward subset regression with independent predictors, situation (1), or with correlated predictors, situation (2), correctly show that the proposed method retains power at any level α, for the variable $V1|V2$. Note that the power in the randomization method is marginally lower than the power under the traditional method (see Figure 2) but this is expected due to the liberal nature of the traditional method. For the significance of the variable $V2|V1$ the distributions of both traditional and randomized p-values were not consistent with the uniform distribution $U(0,1)$.

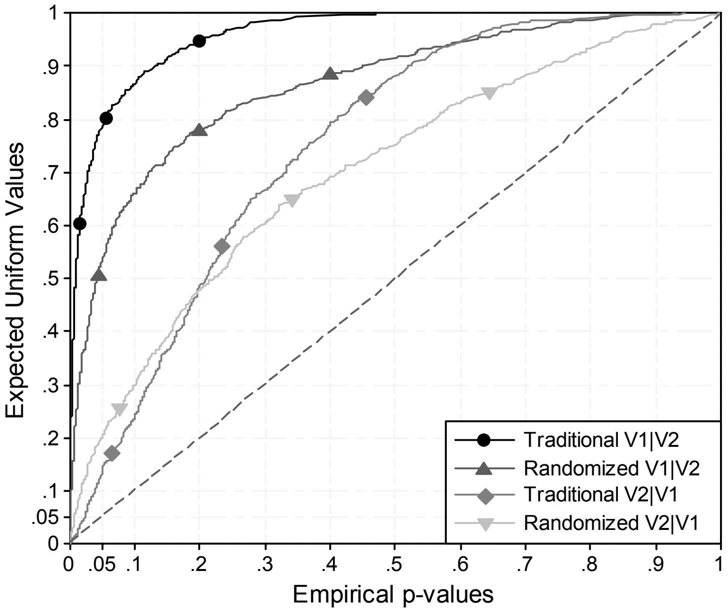

Figure 2: Percentile-Percentile Plot for p-values at Step $k = 2$ for Traditional Approach and for Randomized Approach for $V_{[l,2]}$ under Model B.

6 Effect of the Number of Predictors and Sample Size

The simulation exercise, situation (1), under a true null model (i.e. Model A), was repeated but this time increasing sample size ($I = 30, 60, 90,$ and 120). For each sample size the p-values under the traditional forward modelling approach were not uniformly distributed and the findings mimic those shown in Figure 1. For each sample size the associated p-values for $V_{[l,k]}$ $(l = 1, \dots, k; k = 1, \dots, 4)$ under randomization were calculated. In all cases the empirical distributions of the p-values were consistent with the uniform distribution $U(0,1)$. The differences in the p-values between the traditional forward approach and those obtained under randomization, for $k = 3$ are summarized in Figure 3. In all of these cases the magnitude in the differences show that there is a non-trivial effect which is largely unaffected by sample size when a like for like effect was compared. The same comparable effects were also observed in the simulations under the correlation structure, situation (2).

The above simulation exercise was repeated, but this time altering the number of predictors $(P = 4, 8, 16, 32,$ and $64)$ and keeping the number of cases fixed, $I = 30$, at step $k = 1$. Figure 4 summarizes the distribution of the differences of the p-values under the traditional approach and the randomized approach. In all of these instances the simulations show that the non-uniformity of the p-values under the traditional approach is particularly evident with increasing number of predictors. For the same arrangements, the p-values under randomization were entirely consistent with the uniform distribution $U(0,1)$ and the p-values under the randomization approach were not smaller than the p-values under the traditional approach in any one instance. Note that the discrepancy tends to increase with increasing values of P and this discrepancy is a substantive non-trivial difference.

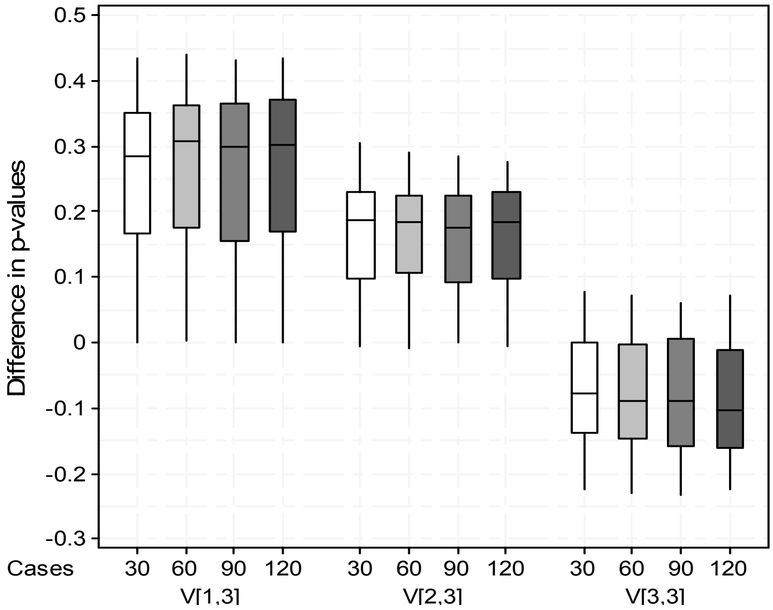

Figure 3: Discrepancy between Randomized and Traditional p-values for Forward Subset Selection at step $k = 3$, under Model A, Situation 1.

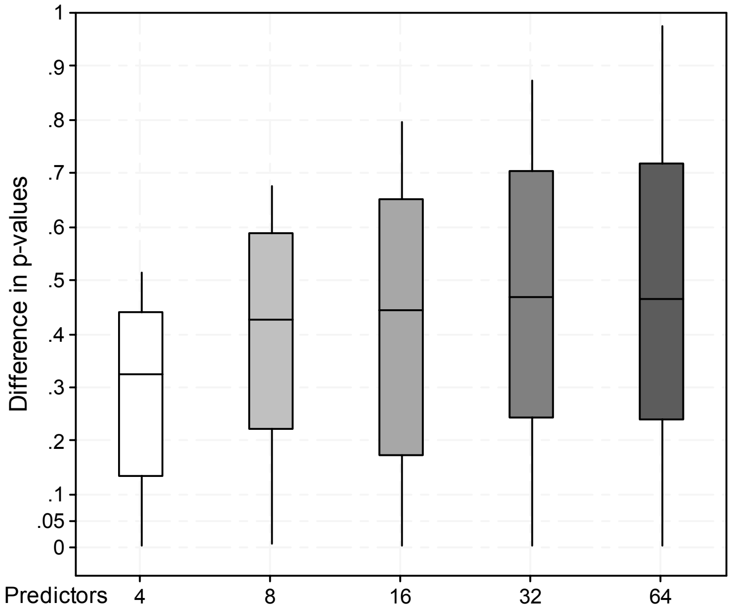

Figure 4: Discrepancy between Randomized and Traditional p-values for Forward Subset Selection at step $k = 1$, under Model A, Situation 1, with Different Number of Predictors.

Consider the simulation exercise reported in Section 4 under Model B, situation 1, with 3 noise variables; where the distribution of the p-values for the second variable $V2|V1$ were not uniformly distributed $U(0,1)$. This time, the simulation exercise was repeated but increasing the number of noise variables under Model B, situation 1, from 3 noise to 9 and 18 noise respectively. Figure 5 summarizes the distribution of the p-values for $V2|V1$ under the traditional approach and the randomization approach. Note that the distribution of the p-values under the traditional approach tends to become increasingly discrepant from the uniform distribution $U(0,1)$, while the discrepancy decreases under randomization.

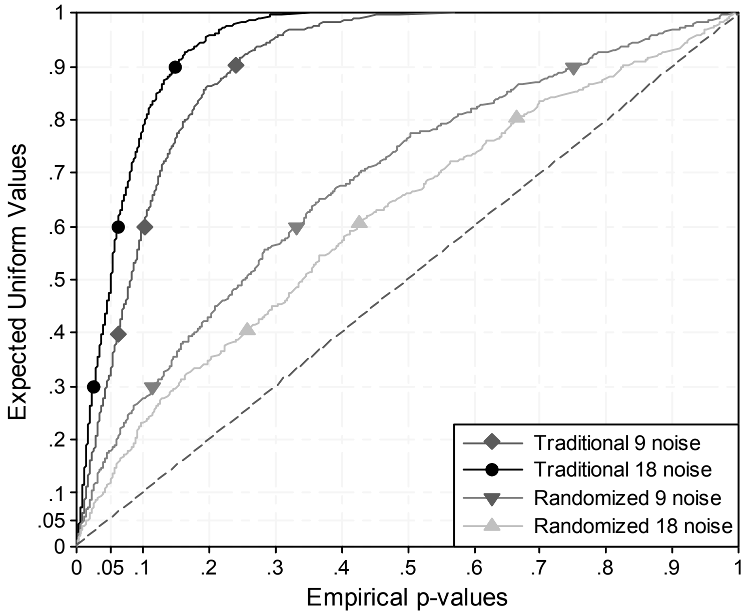

Figure 5: Percentile-Percentile Plot for p-values of $V2|V1$ for Traditional Approach and for Randomized Approach at step $k = 2$, under Model B, with Different Number of Noise Variables.

7 Decision Making for Selecting Subsets

Table 1 summarized the results of $n = 1{,}000$ simulations for each of the two situations (correlated and uncorrelated predictors) under the global null-model, Model A, with $I = 30$ cases and repeated with $I = 120$ cases. Model building was terminated using the $\alpha = 0.05$ stopping criterion. The results are presented in 4×4 cross tabulations where the rows classify the selection using the traditional forward selection approach (missing rows indicates that there are no variables selected) and the columns classify the selection using the randomization approach. The last column labeled "joint selection" represents the final selected model where all variables in the model are deemed to be statistically significant under both the traditional and the randomization approach.

For situation (1) using $I = 30$ cases the traditional forward selection method selected the correct model with no variables 824 times out of the 1000 simulation instances (i.e. at least one Type I error 17.6% of the time). A one variable model was discovered 16.3% of the time and a two variable model 1.3% of the time. Similar percentages are obtained in situation (1) when using $I = 120$ cases. In the correlated cases, situation (2), the traditional forward selection technique similarly showed inflated Type I error rates and the error rates do not seem to substantially depend on sample size.

Contrary to this, for situation (1) using $I = 30$ cases the randomization approach selected the correct model with no variables 941 times out of 1000 (i.e. at least one Type I error at a rate not significantly dissimilar from 0.05). Similar error rates are observed under randomization for the other three cases, demonstrating that error rates using ordered effects under randomization are neither dependent on correlation structure nor on sample size. However, the randomization approach does have a noticeable tendency to incorrectly uncover larger models (e.g. four variable models were obtained under the randomization algorithm but not under the traditional approach). For these reasons we consider the use of joint significance decision rules whereby an effect is to be declared statistically significant if and only if it is significant under both methods. This approach has the tendency of curbing the excessive Type I error rates observed under the traditional approach and to minimize the possibility of relatively larger models that might be uncovered from solely using the randomization technique. The joint selection error rates given in Table 1 show that, using $\alpha = 0.05$, the correct model under a global null-situation is obtained 95% of the time and that increasingly large models occurred with greater rarity, and these findings appear robust to correlation structure and sample size.

Table 1: Final Models Selected by Traditional and Randomization Forward Selection.

			Selection by randomization method						Joint
	Cases	Variables	0	1	2	3	4	Total	Selection
Selection by traditional forward selection		Situation (1): uncorrelated predictors							
	$I = 30$	0	824	0	0	0	0	824	941
		1	111	15	3	2	32	163	52
		2	6	0	0	1	6	13	7
		Total	941	15	3	3	38	1000	1000
	$I = 120$	0	800	0	0	0	0	800	956
		1	141	8	0	0	27	176	37
		2	15	2	0	0	7	24	7
		Total	956	10	0	0	34	1000	1000
		Situation (2): correlated predictors							
	$I = 30$	0	857	0	0	0	0	857	951
		1	89	33	5	3	5	135	46
		2	5	0	0	0	2	7	2
		3	0	0	0	0	1	1	1
		Total	951	33	5	3	8	1000	1000
	$I = 120$	0	863	0	0	0	0	863	941
		1	69	39	8	1	4	121	52
		2	8	0	1	1	5	15	7
		3	1	0	0	0	0	1	0
		Total	941	39	9	2	9	1000	1000

8 Conclusion and Discussion

The foregoing simulations have demonstrated that under a global null hypothesis of no effect that the p-values for variable inclusion under forward selection using standard procedures are not uniformly distributed $U(0,1)$. This applies when the distribution of p-values for the l-th variable ordered on effect size is considered. The extent of the bias is not always in the same direction; the distribution of the p-values entering early into the forward selection solution are stochastically smaller than the uniform distribution $U(0,1)$ (resulting in too many Type I errors) whereas the distribution of the p-values entering later in the forward selection solution are stochastically larger than the uniform distribution $U(0,1)$. These effects do not diminish with increasing sample size and are a function of the forward selection algorithm.

Contrary to this, under a null model the randomization approach retaining the precise correlation structure between sample predictors provides a valid assessment of the statistical significance of the l-th variable at step k ordered on effect size. Under this approach the inference is not on a particular variable X_p at step k but is on $V_{[l,k]}$ at step k. The discrepancy between the traditional p-values and p-values for $V_{[l,k]}$ are not trivial inconsequential differences but could easily be as large as 0.5 (see Figure 3, Figure 4). In common with the standard approach the p-values under randomization are determined using a global null hypothesis: $\beta_p = 0\ (p = 1,2, \dots, P)$ but unlike the traditional approach the randomization method takes into account the data steering part of the forward selection algorithm.

In the non-null situation, the traditional forward selection retains power but is prone to over capitalizing on chance association with non-authentic variables, and the extent of this problem increases with an increasing number of non authentic variables. In the non-null situation, the randomization procedure also retains power and remains prone to including non-authentic variables in a final model but the extent of this bias is smaller in the outlined randomization procedure than in the traditional approach and the extent of the problem diminishes as the proportion of non-authentic variables increases.

The randomization procedure outlined makes an assessment of the significance of variables ordered on their squared partial correlation with the dependent variable. Thus at stage k, an ordered variable $V_{[1,k]}$ has a higher absolute partial correlation than a variable $V_{[2,k]}$ but paradoxically it does not follow that the p-values for $V_{[1,k]}$ under randomization would be smaller than the p-value for $V_{[2,k]}$, which would be the case under the traditional approach.

To partly mitigate this effect we consider a joint decision rule for advancing the forward selection procedure if and only if all variables in a model are statistically significant under both the traditional procedure and under the proposed randomization algorithm. This joint decision making ensures that under a null model the Type I error rate at the first stage is maintained at a pre-determined nominal significance level, it curbs the excessive Type I error rate associated with the traditional forward selection procedure and it reduces the possibility of larger final models that may have otherwise occurred if the randomization procedure had been used by itself.

References

Berk, K. (1978). Comparing subset regression procedures. *Technometrics*, 20(1): 1-6.

Buckland, S., Burnham K., and Augustin, N. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2): 603-618.

Derksen, S., and Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45: 265-282.

Flack, V. and Chang, P. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, 41(1): 84-86.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452): 1304-1308.

Kramer, C. (1957). Simplified computations for multiple regression. *Industrial Quality Control*, 13, 8-11.

Miller, A. (1984). Selection of subsets of regression variables (with discussion). *Journal of the Royal Statistical Society*, *A*, 147(3): 389-425.

Pinsker, I., Kipnis, V. and Grechanovsky, E. (1987). The use of conditional cutoffs in a forward selection procedure. *Communication Statistical Theory and Methods*, 16(8): 2227-2241.