Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 5

11-1-2008

A Randomization Method to Control the Type I Error Rates in Best Subset Regression

Yasser A. Shehata
Productivity and Quality Institute, Yasser.Shehata@live.uwe.ac.uk

Paul White
University of the West of England, paul.white@uwe.ac.uk

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 398-407

1538 – 9472/08/$95.00

A Randomization Method to Control the Type I Error Rates in Best Subset Regression

Yasser A. Shehata
Productivity and Quality Institute

Paul White
University of the West of England

A randomization method for the assessment of statistical significance for best subsets regression is given. The procedure takes into account the number of potential predictors and the inter-dependence between predictors. The approach corrects a non-trivial problem with Type I errors and can be used to assess individual variable significance.

Key words: best subset regression, randomization, Type I error, bias.

Introduction

Subset selection in multiple linear regression is long established: computational algorithms for forward selection techniques date back at least to the 1950's, (see Kramer, 1957), and Canning (1959) gave an example of backward elimination. The use of subset selection techniques is widespread and continuing. George (2000) wrote "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 use of automated computer techniques for model building is rife. Some researchers use automated search algorithms as a data mining exercise (Lovell, 1983), examining a research question by collecting data on virtually every variable that could possibly be related to the phenomenon under investigation and attempting to obtain a parsimonious model based on patterns in sample data. In recognition of this type of problem Larzelere and Mulaik (1977) suggested basing inferences on the total number of potential predictors rather than the number of predictors in a given subset.

It is commonly argued that a purpose of automated selection techniques is to obtain a simple, high-quality representation of the phenomenon under investigation. This is accomplished by not including potential predictors deemed to be uninformative in a final model. Models based on smaller numbers of predictor variables are comparatively easier to understand and it is hoped that a parsimonious model will give greater insight into the underlying processes that generated the data. In some instances smaller subsets may lead to greater economy (Derksen & Keselman, 1992).

Problems relating to variable selection from using backward elimination, forward selection, best subset regression and other automated model building techniques are well documented in the context of multiple linear regression. Investigations have generally been through simulation work in which the theoretical underpinning model assumptions are satisfied and any deviation between simulation results and anticipated theoretical results is therefore attributable to the variable selection technique. For instance, the simulation work of Derksen &

Yasser A. Shehata is Lecturer at the Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt. Email: Yasser.Shehata@live.uwe.ac.uk. Paul White is Senior Lecturer and a member of the Applied Statistics and Quantitative Methods Consultancy in the Department of Mathematics & Statistics, Bristol, UK. Email: Paul.White@uwe.ac.uk.

Keselman (1992) gave broad the conclusions that automated selection techniques overly capitalize on false associations between potential predictors and the criterion variable with too many purely random (noise) variables being wrongly classified as authentic (true) predictors. The inclusion of noise variables in a final model necessarily implies a model misspecification or misidentification and incorrect inferences are drawn. Derksen & Keselman (1992) additionally found that the incidence with which noise and authentic variables find, or do not find, their way into a final model depends upon the degree of correlation between predictor variables. As such, it would seem that controlling the error rate may require a solution which explicitly utilizes within sample correlation information.

Hurvich & Tsai (1990) pointed out that, if a model is not fully pre-specified and, if a model selection technique is used, then the number of regression parameters is a random variable. Moreover, once a model has been decided upon by some technique, the model estimation and the associated hypothesis tests usually proceed on the assumption that the data driven and technique selected model is the true model. In other words, the data is analyzed "as though they were a fresh data set generated by the selected model" (Hurvich & Tsai, 1990, p. 214). Under these conditions, as pointed out by Miller (1984), the regression estimators may be biased and standard hypothesis tests may not be valid.

Automated model building techniques, such as stepwise regression, proceed on the basis of performing many statistical tests and do so in instances whereby the hypothesis test procedure may not be valid. Multiplicity of testing contributes to model selection problems. In the context of stepwise regression Derksen & Keselman (1992) wrote "when many tests of significance are computed in a given experiment, the probability of making at least one Type I error in the set of tests, that is, the maximum familywise Type I error rate (MFWER), is far in excess of the probability associated with any one of the tests" (p. 269). In subset selection there are a potentially large number of statistical tests to be performed to drive the algorithms. The number of such tests is not known in advance and simple Bonferroni corrections may be too liberal in correcting this problem, especially when potential predictors are not orthogonal. Paradoxically, others have suggested that a more liberal approach is appropriate. Bendel & Afifi (1977) advocated the use of nominal significance levels between $\alpha = 0.15$ and $\alpha = 0.25$ in forward selection so as to include all authentic variables at the expense of an increased risk of including additional noise variables in a final model.

The all subsets approach searches through all possible subsets for each subset size of $1,2,.....,J$ and best subsets chooses the one that has the best summary statistics for a given subset size. A possible best summary statistic is the R^2 statistic (the coefficient of determination). An advantage of the best subsets and all subsets approach over sequential procedures is that this approach, by definition, will not miss finding the best fitting subset of any given size. Indeed, Mantel (1970) pointed out, and gave instances and explanations that a multivariate combination of variables might produce the best fit, but these multivariate combinations might not be identified by sequential procedures. Further, Kuk (1984) pointed out a relative weakness of sequential procedures in that "they lead to a single subset of variables and do not suggest alternative good subsets" unlike all subsets and further points out that sequential procedures have "the possibility of premature termination" (Kuk, 1984, p. 587). Identification of best subsets need not necessarily be computationally burdensome as the identification of the best subset does not require the calculation of all possible subsets (Furnival & Wilson, 1974).

The above provides a strong rationale for considering best subsets regression. The standard inferential approach for best subsets regression has problems arising from using standard hypothesis tests based on a global null hypothesis of no effect for a model determined by sample data. Motivated by the stance of Larzelere & Mulaik (1977) the use of randomization to control Type I error rates in best subsets regression is considered, and the approach takes into account the total number of predictors under consideration. Derksen & Keselman (1992) concluded that the extent of

the problem with automated techniques depends upon the degree of correlation between predictor variables. The use of randomization permits the correlation structure between potential predictor variables to be accounted for. The approach adopted is to compute *p*-values for overall model significance and for each variable under a global null model (as per standard approaches) but which will correct the bias associated with the procedural aspects of best subsets regression. Randomization additionally permits a like-for-like comparison for individual variables that comprise a best subset solution; topics which are expanded in this article.

A brief overview of the traditional least squares approach to determine overall model significance of a best subset regression solution in addition to the individual significance of the variables that comprise the model is first given. Next, a randomization approach that empirically estimates overall and individual significance of best subset regression is described. Descriptions of two models are given, namely a global null-model and a non-null model. These two models, under certain conditions, are used to compare the performance of the randomization algorithm with the traditional approach. Results of the simulation, effects of number of predictors and effects of sample size are provided. The discussion addresses issues concerning the paradoxical problems associated with judging inference in best subsets regression.

Methodology

Best Subsets Regression

Consider the classic linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_J X_J + \varepsilon \quad (1)$$

where Y is the dependent variable with J predictors $X_1, X_2, ..., X_J$ and where ε denotes a normally distributed random variable. Let $y_i, x_{1i}, x_{2i}, ..., x_{Ji}, \ (i = 1, 2, ..., I)$ denote I independent cases generated from the above model.

In best subsets regression, the best subset of size j is the subset of j predictor variables that maximizes the within sample prediction of the dependent variable, y, in a linear least squares regression. This is the percentage of variation in y that is accounted for by a regression equation is the usual R^2 statistic. In the following, R_j^2 will be used to denote the R^2 statistic for the best subset of size j. Overall significance of the best subset of size j is judged using the standard F statistic, $F = S_R^2 / S_E^2$ where S_R^2 is the mean square due to regression, S_E^2 is the mean square error and overall model significance is judged by making reference to the F distribution with $(\upsilon_1, \upsilon_2) = (j, I - j - 1)$ degrees of freedom.

The relative magnitude of the observed value of the F statistic is quantified by the *p*-value and contemporary practice is to declare a statistically significant subset of predictors whenever $p < 0.05$. In addition, let S_p^2 denote the change in the error sum of squares for deleting a variable X_p from a regression model. An assessment of the statistical significance of X_p in the model is made by referring $F = S_p^2 / S_E^2$ to the F distribution with degrees of freedom $(\upsilon_1, \upsilon_2) = (1, I - j - 1)$. For a detailed explanation of best subsets of regression see Draper & Smith (1981, p. 303).

If the potential predictor variables $X_j, (j = 1, 2, ..., J)$, are noise variables, i.e. unrelated to Y in as much as $\beta_j = 0, (j = 1, 2, ..., J)$, then the *p*-values for judging overall model significance for any subset of size j, should be uniformly distributed $U(0, 1)$. Thus, if a researcher works at the α significance level and, if none of the potential predictor variables are related to Y, then a Type I error in assessing significance of the overall best subset model should only be made $\alpha\%$ of the time for any value $\alpha \in (0,1)$. Arguably, the same requirement should also apply to individual predictor variables. An alternative procedure for assessing the overall significance of any best subset of size j and for assessing the statistical significance of each variable included in the best subset model is proposed. This alternative procedure, a randomization method, does not make explicit use of the properties of

the *F* distribution. Ordering the variables that comprise a best subset solution in terms of their individual *F* values is also considered along with deriving an estimate of their *p*-value by considering similarly ordered *F* values under randomization.

Randomization

Consider sample data $y_i, x_{1i}, x_{2i}, ..., x_{Ji}$, $(i = 1, 2, ..., I)$, and let R_j^2 denote the coefficient of determination for the best subset of size $j, (j = 1, 2, ..., J)$. Next consider where the order of cases for the predictor variables in the data is randomly permuted but with the response variable held fixed at $y_i, x_{1i}, x_{2i}, ..., x_{Ji} \rightarrow y_i, x_{1k}, x_{2k}, ..., x_{Jk}$. 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.

Best subsets regression can be performed on the newly created randomized data set. Let S_j^2 denote the coefficient of determination for the best subset of size $j, (j = 1, 2, ..., J)$ for the randomized data set. If for subset $j, S_j^2 > R_j^2$, then the randomized chance solution may be viewed as having better within sample predictability than the observed data.

For any given data set many permutations of the original data set may be generated by taking another random permutation. In what follows the proportion of instances that $S_j^2 > R_j^2$ is estimated through simulation. This estimate is taken to be an estimate of the *p*-value for determining the statistical significance of R_j^2 for any subset of size *j*. For a given data set, an increase in the number of random permutations will serve to increase the accuracy of the estimated value.

The above procedure may be summarized as follows:

For given data set and for a subset of size *j*:

1. Determine the best subset of predictors of size *j* and record the coefficient of determination R_j^2
2. Set KOUNT = 0
3. DO n = 1 TO N
   a. Randomly permute $x_{1i}, x_{2i}, ..., x_{Ji}$ independently of y_i i.e. $y_i, x_{1i}, x_{2i}, ..., x_{Ji} \rightarrow y_i, x_{1k}, x_{2k}, ..., x_{Jk}$
   b. For the newly created fake data set determine the best subset of size *j* and record the coefficient of determination S_j^2
   c. If $S_j^2 > R_j^2$ then KOUNT = KOUNT+1
4. ENDDO
5. Estimated *p*-value = KOUNT/N

The counting process effectively estimates rank position of the original solution in relation to randomization solutions. Under the randomization process all permutations are equally likely. Likewise if the original predictors are generated under a system whereby none of them are related to the outcome then the observed value of R_j^2 is just a likely to be as large as any value of S_j^2 obtained from random permutation.

In a similar way for best subset of size *j*, consider the *F*-values for each predictor variable arranged in order, $F_{(1)} > F_{(2)} > ..... > F_{(j)}$. The *F*-values from a random permutation may be ordered in a similar way, i.e. $F_{(1)}^* > F_{(2)}^* > ..... > F_{(j)}^*$. The proportion of times $F_{(p)}^* > F_{(p)}$ provides an estimate of the *p*-value of the *p*-th ordered variable in the observed best subset solution.

Simulation Design

For a specific application consider the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (2)$$

To illustrate the properties of the proposed technique, four specific parameter settings

(referred to in the following as Model A and Model B) with two different correlation structures have been considered.

Model A is a genuine null model with $\beta_0 = 1$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ so that all proposed predictors are noise variables and are unrelated to the outcome *Y*. For Model B consider $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = \beta_3 = \beta_4 = 0$ (i.e., one authentic variable and three noise variables).

In the following simulations each model is considered with potential predictor variables being (i) Case 1, stochastically independent in which their correlation matrix is the identity matrix, and (ii) Case 2, strongly correlated 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 between X_l and X_m.

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. Simulations herein are reported based on *I* = 30 cases per simulation instance and increasing sample size and increasing the number of potential predictors are considered.

Results

Figure 1 is a percentile-percentile plot of the *p*-values obtained from implementing the aforementioned algorithm for step *j* = 1, 2, 3 in best subsets regression for Model A with potential predictor variables being stochastically independent. The vertical axis denotes the theoretical percentiles of the uniform distribution *U*(0, 1) and the horizontal axis represents the empirically derived percentiles based on 500 simulations with each simulation based on 1,000 randomization instances. Note that the *p*-values based on the traditional method are systematically smaller than required, indicating that the true Type I error rate for overall model significance is greater than any pre-chosen nominal significance level α. By contrast the estimated *p*-values based on the randomization algorithm have an empirical distribution that is entirely consistent with the uniform distribution *U*(0, 1) for any subset of size 1, 2, or 3 out of 4 predictors.

Under Model A, qualitatively similar results are obtained for *j* = 1, 2, 3 for potential predictors being correlated, Case 2. For *j* = 4 there is no subset selection under the simulations and in these cases both the traditional method and the randomization method have *p*-values uniformly distributed *U*(0, 1).

Simulations under Model B for step *j* = 1, 2 in best subsets regression with independent predictors, Case 1, or with correlated predictors, Case 2, correctly show that the proposed method retains power at any level of α; the power 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.

Once overall model significance has been assessed, a normal practice is to assess the individual significance of each variable alone. Figure 3 is a percentile-percentile plot of the *p*-values for the variables that comprise the best subset of size *j* = 3 of 4 under Model A, Case 1. In this instance the three variables included in the model have been ordered according to their *F*-values. The traditionally computed *p*-value for the variable with the largest *F*-value is typically too small when judged against the uniform distribution, *U*(0, 1). Contrary, for the variable with the smallest *F*-value the *p*-values calculated using the standard method are typically too large when judged against the uniform distribution, *U*(0, 1). By contrast, the *p*-value under the randomization method, for all ordered effects, is entirely consistent with the uniform distribution *U*(0, 1).

Qualitatively similar results are obtained for Model A but for potential predictors being correlated, Case 2.

Simulations under a true null model (i.e. with all potential predictors being noise variables), for *J* = 4, 8, 16, 32, 64 keeping the number of cases fixed, *I* = 30, have been performed. In all of these cases the simulations show that the *p*-value for overall subset

Figure 1: Percentile – Percentile plot for *p*-values for overall significance for best subset of size *j* = 1, 2, 3 from 4 independent predictors, Model A.

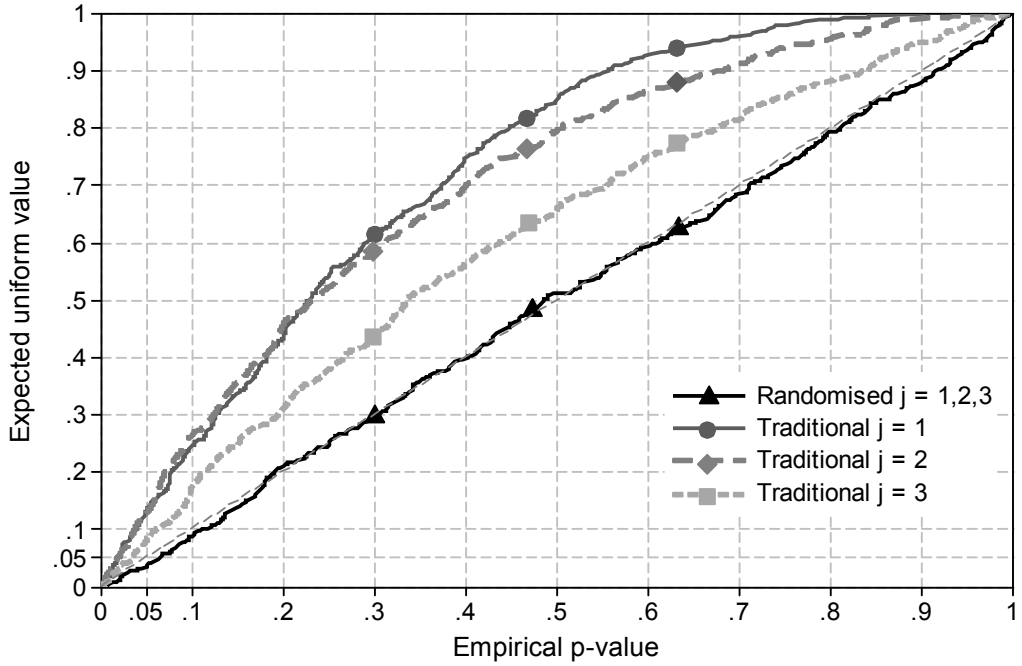


Figure 2: Percentile – Percentile plot for *p*-values for best subset of size *j* = 1, 2 from 4 independent predictors, Model B.

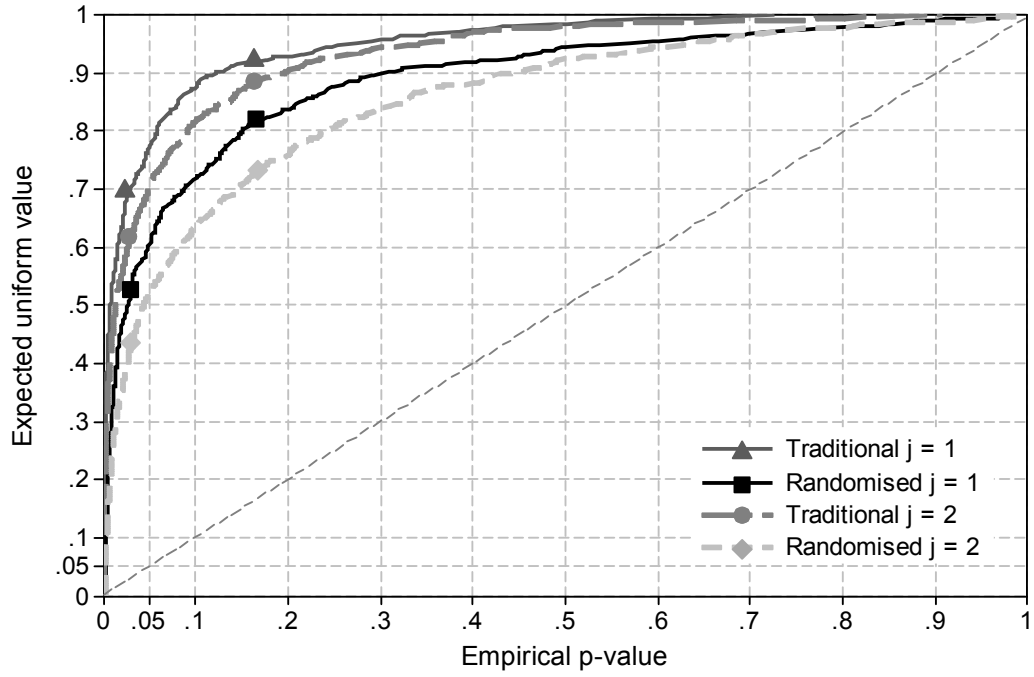
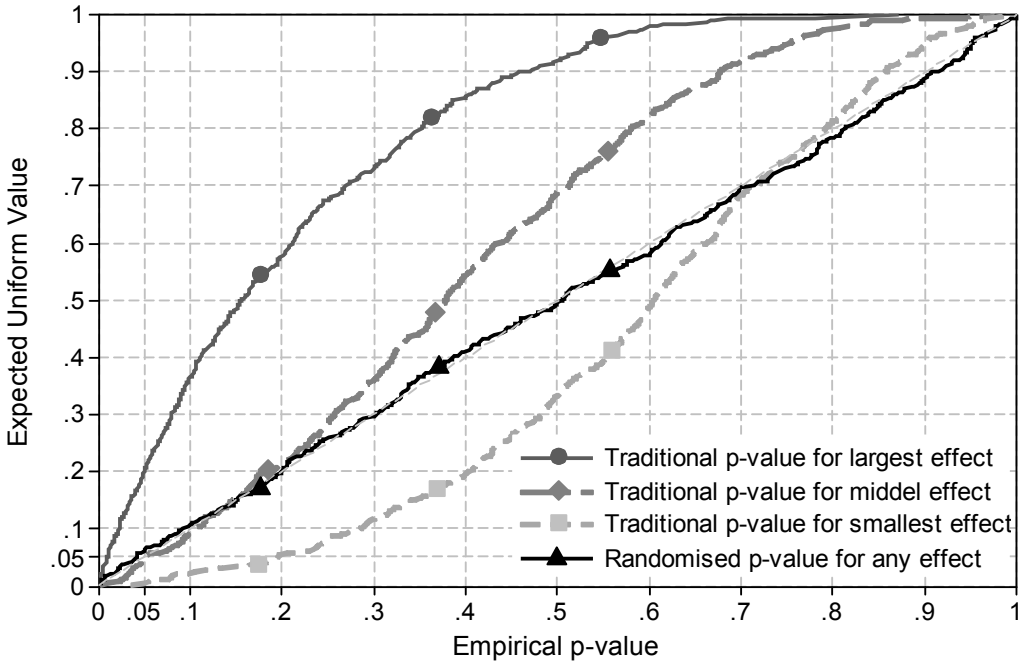


Figure 3: Percentile – Percentile plot for *p*-values for each variable in a best subset of size j = 3 from 4 independent predictors when the effect size is order by magnitude, Model A.



significance using the proposed randomization method is uniformly distributed $U(0, 1)$.

In every simulation instance the estimated *p*-value in the randomization method for overall model significance was not less than the *p*-value under the traditional method. The distribution of the differences for $j = 1$ and $J = 4$, 8, 16, 32, 64 is summarized in Figure 4. Note that the discrepancy tends to increase with increasing values of J and that this discrepancy is a substantive non-trivial difference.

Simulations under a true null model (i.e., with all potential predictors being noise variables), for J = 4, 8, 16, 32, 64, but with different sample sizes, I = 30, 60, 90, 120 have been performed. In all of these cases the simulations show that the distribution of *p*-value for overall subset significance using the proposed randomization method is uniform $U(0, 1)$. In every simulation instance the estimated *p*-value using the randomization method is not less than the *p*-value under the traditional method. Figure 5 summarizes the extent of the differences.

Figure 4: Discrepancy between randomized and traditional *p*-values for best subset of size $j = 1$ with $I = 30$ and different number of predictors.

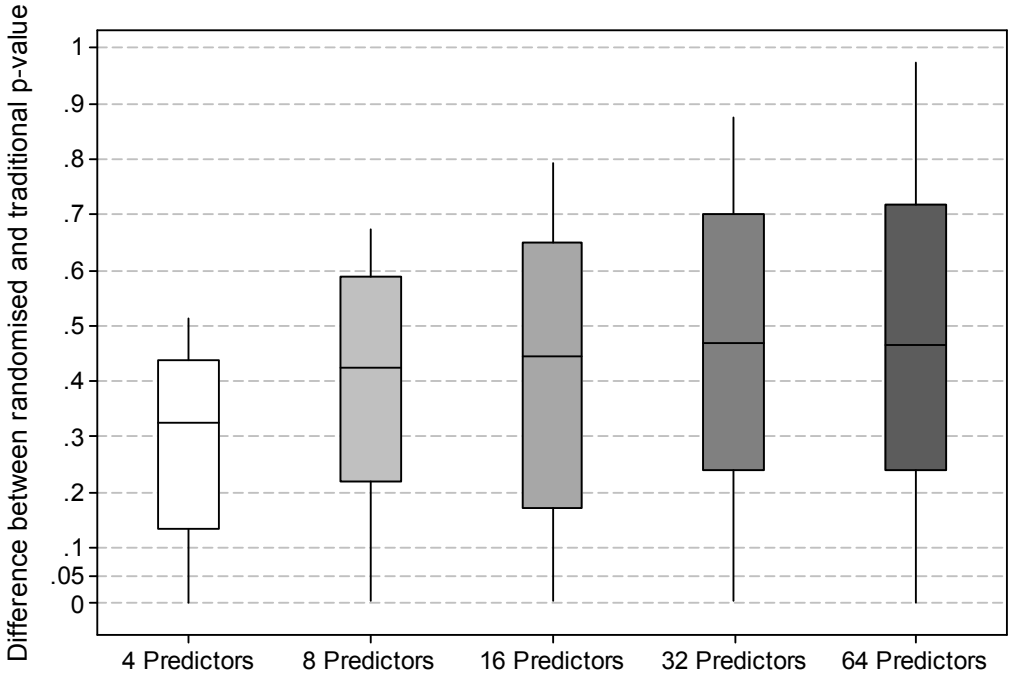
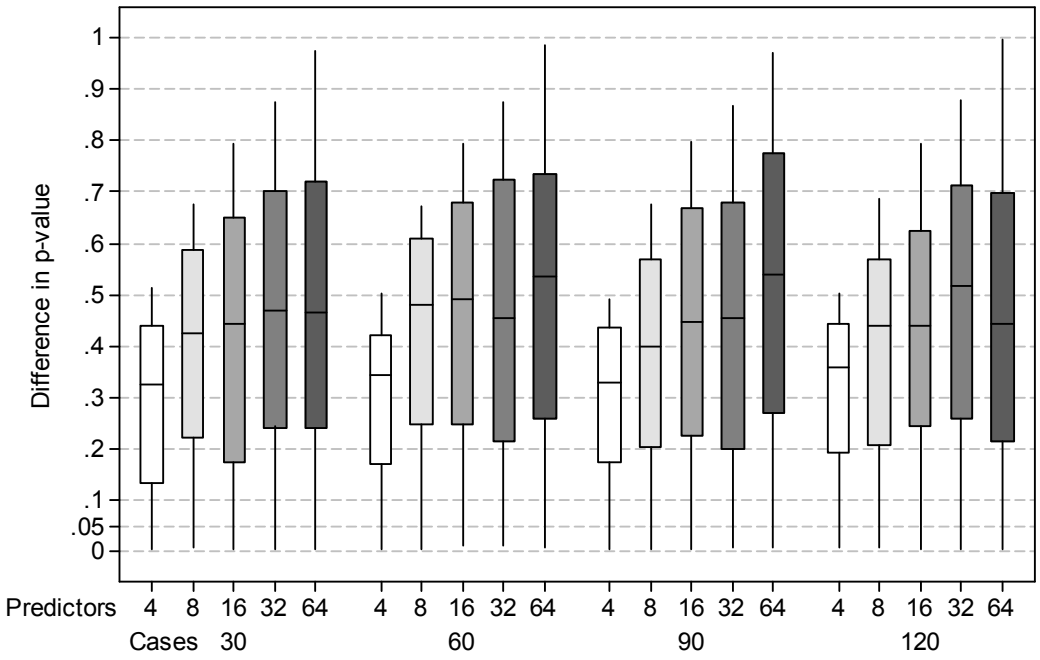


Figure 5: Distribution of the difference in *p*-values for overall model significance under both the randomization and the traditional methods for Model A for subset of size j =1.



Conclusion

A computer based heuristic that uses randomization has been described. The algorithm allows control of Type I error rate for the overall statistical significance of a best subsets regression model and control for the variables that comprise the model based on their relative order. This randomization algorithm permits the Type I error rate to be controlled at any pre-determined nominal significance level, α. The data sets created under the randomization procedure, each precisely retained the correlation structure observed in the original data and, as such, the approach takes into account the data set dependent problems that arise due to the correlation structure between potential predictor variables (see Derksen & Keselman, 1992). For the *j*-th best subset the procedure produces *p*-values indirectly based on the number of potential predictor variables (*J*) rather than the number of predictor variables in a given subset (*j*) and, as such, retains some similarity with the stance of Larzelere & Mulaik (1977). Their approach, however, does not take into account the correlation structure between potential predictor variables. By contrast, the algorithm outlined in this article establishes the *p*-value for overall model significance based on the effective number of predictors. For example consider J potential predictors, and consider an extreme case whereby $J-1$ of the predictors are mutually orthogonal but the other predictor is perfectly correlated with one of the other predictors in the orthogonal set. In this extreme case the number of predictors is J but the number of effective predictors is $J-1$.

The simulation work demonstrates that the randomization algorithm corrects a non-trivial problem. This correction also applies in those particularly problematic cases whereby the number of predictors exceeds the number of cases (subject to subset size *j* being less than sample size *I*).

Significance tests in classical least squares regression are based on the assumption that the underpinning error terms are independent, identically distributed normal random variables. When these assumptions are satisfied the *p*-value for overall model significance for a best subsets regression of size *j* still displays a bias. By contrast, the corresponding *p*-value estimation using the randomization algorithm does not suffer from this bias. In practice the underpinning normality assumptions are likely to be violated to some extent, and these violations may lead to additional biases in the estimated *p*-values for overall model significance in a best subsets regression using the standard approach. The randomization approach is based on the sample data and the estimation of the *p*-value does not explicitly rely upon distributional assumptions. Indeed, the algorithm is not peculiar to ordinary least squares regression and could be applied to other classes of model, including those models that rely upon inferential tests of significance based upon large sample asymptotic theory (e.g. binary logistic regression).

The approach for assessing individual significance of variables that comprise a final best subset is to consider a rank ordering on the variables in the model according to the value of their corresponding *F* statistic. This imposition of an ordering allows for a fair comparison with similarly ordered variables in the randomized solutions. It is recognized that this may produce a seemingly paradoxical outcome in some situations. For instance, and for simplicity of exposition, consider a two variable subset $j=2$ with a variable, X_1 with *F*-value $F_{(1)}$ and a variable, X_2, with *F*-value $F_{(2)}$. Without loss of generality assume $F_{(1)}>F_{(2)}$. In evaluating the statistical significance of X_1, the value $F_{(1)}$ will be compared against similarly ordered values $F^*_{(1)}$ and the value $F_{(2)}$ will be compared with similarly ordered values $F^*_{(2)}$. No condition is imposed to ensure that the proportion of times $F^*_{(1)} > F_{(1)}$ is less than the proportion of times $F^*_{(2)} > F_{(2)}$. However it should be borne in mind that X_1 and X_2 were not specified in advance; rather the significance tests alluded to are tests of significance for the variable with the largest value $F_{(1)}$ and for the variable with the second largest value $F_{(2)}$. In practice, interest would

focus on those final solutions where all variables in the model met some pre-defined nominal level of significance (e.g. $\alpha = 0.05$).

A motivation behind this research was to help develop a sound methodological process to assist researchers in constructing valid and good initial models in exploratory research. However, the use of automated techniques is not in itself a substitute for quality of thought in determining what may be a good predictor of an outcome variable. An understanding of the procedural aspects involved in assessing statistical significance through the use of randomization may have an added benefit of focusing a researcher to determine seemingly good predictors at the outset rather than a researcher collecting data on all conceivable predictors and using these without penalty as per procedures currently offered by standard statistical software.

References

Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward 'stepwise' regression. *Journal of the American Statistical Association*, *72*, 46-53.

Canning, F. L. (1959). Estimating load requirements in a job shop. *Journal of Industrial Engineering*, *10*, 447.

Derksen, S., & Keselman, H. J. (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.

Draper, N. & Smith, H. (1981). *Applied regression analysis*. John Wiley & Sons: New York, NY.

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, *16*, 499-511.

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*, 1304-1307.

Hurvich, C. M., & Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44, 3*, 214-217.

Kramer, C. Y. (1957). Simplified computations for multiple regression. *Industrial Quality Control*, *13, 8*, p8.

Kuk, A. (1984). All subset regression in a proportional hazards model. *Biometrika*, *71, 3,* 587-592.

Larzelere, R. E., & Mulaik, S. A. (1977). Single-sample tests for many correlations. *Psychological Bulletin*, *84*, 557 – 569.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, *65*, 1-12.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, *12, 3*, 621-625.

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