A COMPUTATIONAL APPROACH TO CORRECTLY ASSESS SIGNIFICANCE IN BEST SUBSET REGRESSION

Yasser. A. Shehata[1] and Paul. White[2]

[1]Productivity and Quality Institute, Arab Academy for Science and Technology, Alexandria, Egypt
Yasser.Shehata@uwe.ac.uk

[2]University of the West of England, School of Mathematical Sciences, Bristol, UK
Paul.White@uwe.ac.uk

ABSTRACT

Best subsets regression is often used to identify a good regression model. The standard approach to assess statistical significance for a best subsets regression model is flawed. A computationally intensive randomization algorithm which corrects the problem is outlined and implemented. Simulation studies show that this procedure corrects a non-trivial problem that exists independent of sample size and is a procedure that is robust to the presence of influential observations. This procedure leads to a simple decision rule even with correlated predictors unlike the use of a single probe. The proposed method is shown to retain power in a non-null situation.

Keywords: *Best subset regression, randomization, probe variable, Type I error, bias.*

1. INTRODUCTION

A first stage in the development of a good predictive model or a good classification rule is the identification of potentially useful predictor variables based on domain knowledge. The general type of model to be developed also needs to be defined. Depending on the circumstances the type of model to be considered could, for instance, be a linear regression model, or a logistic regression model, or a regression tree or a neural network.

In exploratory model building the selection of appropriate variables for inclusion in a final model is often done algorithmically. Thus for instance, algorithms such as backward elimination, forward selection or best subsets are routinely employed to develop regression models (see [5]). The motivation behind the development of the cited algorithms is to have a procedure that will identify a good subset of predictor variables. In this sense the ideas of variable selection and subset selection become synonymous. The use of these algorithms in regression problems is widespread even though their use is known to be problematic. The extent of the use of the algorithmic approach in model building is aptly summarized by George [4], who writes "*The problem of variable selection is one of the most pervasive model selection problems in statistical applications. 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*".

Variable selection problems 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. In the main, investigations have 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 and Keselman [2] give broad 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 and incorrect inferences are drawn.

Derksen and Keselman [2] also concluded that the inclusion of noise variables in a model can result in the failure to classify genuine (authentic) variables as being genuine predictors of the criterion. Thus, well established automated techniques can paradoxically inflate the probability of Type I errors and in some cases result in a loss of power. Moreover, the conclusions drawn by Derksen and Keselman [2] indicate that "*the degree of correlation between predictor variables affected the frequency with which authentic variables found their way into the model*". Accordingly the rate at which Type I errors occur is quite problem dependent and there is no simple mechanism for controlling this error rate.

The over capitalization on false associations leads to overfitting and gives rise to overly optimistic within sample estimates of goodness-of-fit and overly optimistic predictive ability which is not replicated on new data from the same population. Best subset regression solutions are based on the overall within sample maximization of the goodness-of-fit statistic, and these "best subset" solutions necessarily show the greatest upward bias in the estimation of the population coefficient [6]. This problem is compounded when the number of potential predictor variables J increases relative to the number of cases I [6]. For these reasons we consider an alternative technique to correctly quantify the Type I error rate in assessing overall model significance for best subset regression solutions and further show that the correction under the proposed method is a non-trivial correction especially in cases where $J > I$.

In Section 2 we give a brief overview of the traditional least squares approach to determine overall significance of a best subset regression solution. In Section 3 we outline an alternative randomization approach. In Section 4 a description of a series of models is given that will be 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 given in Sections 5 and 6 respectively. We additionally discuss the robustness of the proposed procedure to influential observations (Section 7). Our discussion also casts doubt on the use of a single probe (Section 8) variable for assessing overall model significance.

2. BEST SUBSETS REGRESSION

Consider the classic linear regression model

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

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

In best subsets regression, the best subset of size j is that subset of j predictor variables that maximizes the within sample prediction of the dependent variable, y, in a linear least squares regression. The percentage of variation in y that is accounted for by a regression equation is the usual R^2 statistic, known as the coefficient of determination. 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 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 with value of $p < 0.05$ traditionally taken to indicate an overall statistically significant subset of predictors. For a more detailed explanation of best subsets of regression see [5].

If the potential predictor variables X_j, $(j = 1, \cdots\cdots, J)$, are noise variables i.e. unrelated to Y in as much as $\beta_j = 0$, $(j = 1, \cdots\cdots, J)$, then the p-values for judging overall model significance, for any subset of size j, should be uniformly distributed on (0, 1). That is to say, 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 $\alpha \in (0,1)$. We propose an alternative procedure for assessing the overall significance of any best subset of size j. This alternative procedure, a randomization or "fake variable" method, does not make explicit reference to the F distribution.

3. FAKE VARIABLE METHOD

Reconsider the sample data $y_i, x_{1i}, x_{2i}, \ldots\ldots, x_{Ji}$, $(i = 1, \cdots\cdots, I)$ and let R_j^2 denote the coefficient of determination for the best subset of size j, $(j = 1, \cdots\cdots, J)$. 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. $(y_i, x_{1i}, x_{2i}, \cdots\cdots, x_{Ji}) \rightarrow (y_i, x_{1k}, x_{2k}, \cdots\cdots, x_{Jk})$. Note that this random permutation of predictor records ensures that the sample correlation structure between the predictors in the real data set is precisely preserved in the newly created randomized data set (also known as the "fake data set"). The random permutation also ensures that the predictor variables in the fake 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 fake data set. Let S_j^2 denote the coefficient of determination for the best subset of size j, $(j = 1, \cdots\cdots, J)$ for the fake data set. If for subset j $S_j^2 > R_j^2$ then the fake "chance" solution may be viewed as having better within sample predictability than the observed data.

Naturally, for any given data set many instances of a fake data set may be generated simply 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 increase the accuracy of the estimated value. The above procedure may be summarized as follows:

```
For given data 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 $(x_{1i}, x_{2i}, \ldots, x_{Ji})$ independently of $y_i$ i.e. $(y_i, x_{1i}, x_{2i}, \cdots, x_{Ji}) \rightarrow$
      $(y_i, x_{1k}, x_{2k}, \cdots, x_{Jk})$
   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
```

4. Simulation Design

For a specific application consider the model

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

To illustrate the properties of the proposed technique, four specific parameter settings (referred to in the following as Model A, Model B, Model C, and Model D) with two different correlation structures have been considered.

Model A is a genuine null model with $\beta_0 = 1$ and with $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ i.e. all proposed predictors are in fact noise variables and are unrelated to the outcome Y. For Model B we consider $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = \beta_3 = \beta_4 = 0$ (i.e. one authentic variable and three noise variables). For Model C we consider $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.2$, and $\beta_3 = \beta_4 = 0$. For Model D we consider $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.2$, $\beta_3 = 0.1$, and $\beta_4 = 0$.

In the following simulations each model is considered with potential predictor variables being (1) stochastically independent in which their correlation matrix is the identity matrix, and (2) strongly correlated with elements of the correlation matrix being $\rho(X_1, X_2) = 0.708$, $\rho(X_1, X_3) = 0.802$, $\rho(X_1, X_4) = -0.655$, $\rho(X_2, X_3) = 0.757$, $\rho(X_2, X_4) = -0.582$, and $\rho(X_3, X_4) = -0.593$ 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 linear regression models are satisfied. 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 predictors

5. Simulation Results

Fig. 1 is a percentile-percentile plot of the *p*-values obtained from implementing the aforementioned algorithm for step *j* = 1 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 (0, 1) and the horizontal axis represents the empirically derived percentiles based on 500 simulations. 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. In contrast the estimated *p*-values based on the fake variable data set have an empirical distribution that is entirely consistent with the uniform distribution (0, 1).

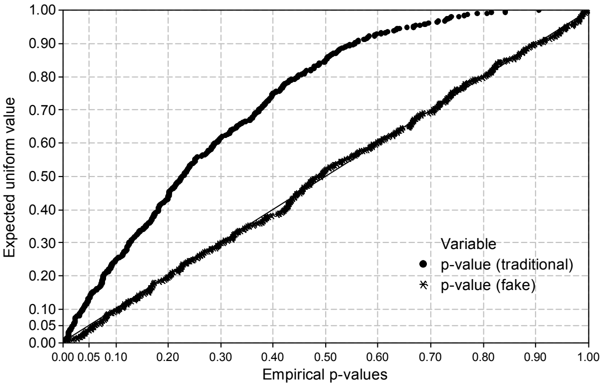

Fig. 1. Percentile – Percentile plot for *p*-values for best subset of size 1 from 4 independent predictors, Model A.

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

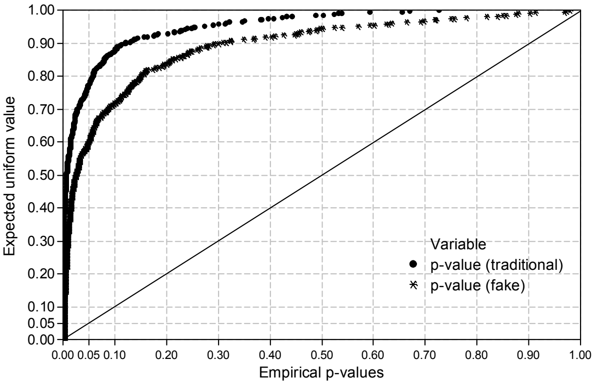

Fig. 2. Percentile – Percentile plot for *p*-values for best subset of size 1 from 4 independent predictors, Model B.

Simulations under Model B, C, and D 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 Fig. 2.) but this is expected due to the liberal nature of the traditional method as evidenced in Fig. 1.

6. Effect Of The Number Of Predictors And Sample Size

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 subset significance using the proposed fake variable method is uniformly distributed on (0, 1).

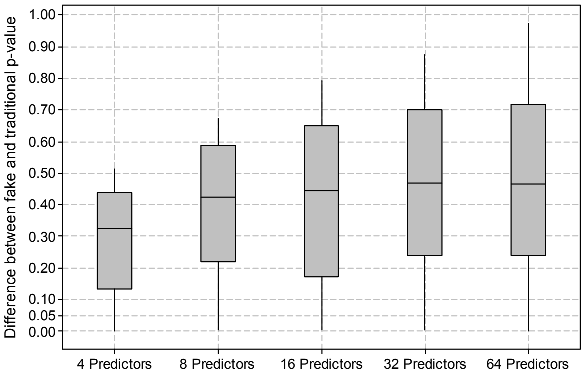

Fig. 3. Discrepancy between fake and traditional p-values for best subset of size 1 with different numbers of predictors.

In each and every simulation instance the estimated p-value in the fake variable method is 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 Fig. 3. 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 subset significance using the proposed randomized method is uniform on (0, 1). In each and every simulation instance the estimated p-value in the fake variable method is not less than the p-value under the traditional method. Fig 4 summarizes the extent of the differences.

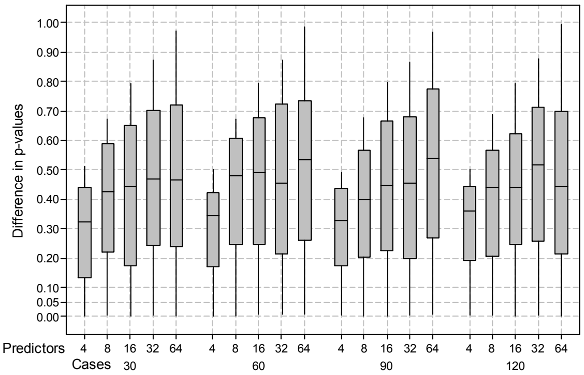

Fig. 4. Distribution of the difference in p-values under the fake variable method and the traditional method for Model A.

7. Effect Of Outliers And Influential Observations

The simulations referred to in Section 5 have been repeated but with the inclusion of (i) a single outlying observation, (ii) a single influential observation, and then (iii) with a single influential outlying observation.

The introduction of a single observation with high leverage ($x_{1,30}, x_{2,30}, x_{3,30}, x_{4,30}$) = (4, 4, 4, 4) under Model A (with either stochastically independent predictors or correlated predictors) did not grossly affect the distribution of p-values under the proposed randomization method. Likewise the introduction of a single observation with high leverage, ($x_{1,30}, x_{2,30}, x_{3,30}, x_{4,30}$) = (4, 4, 4, 4), under Model B and with a simulated response consistent with Model B did not affect the distribution of p-values under the proposed randomization method irrespective of the correlation structure between predictors. Similarly the introduction of a single observation with high leverage, ($x_{1,30}, x_{2,30}, x_{3,30}, x_{4,30}$) = (4, 4, 4, 4), under Model B but with a simulated response not consistent with Model B did not materially affect the distribution of p-values under the proposed randomization method.

8. A Probe Variable

For comparative purposes, simulations were performed under Model A with the inclusion of a single probe variable. The simulations proceeded along the following lines:

1. generate a data set ($y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}$) under Model A
2. generate realizations of a random variable Z where Z is stochastically independent of Y and independent of X_1, X_2, $X_{3,}$ and $X_{4.}$
3. include the values of Z in the data set to form an augmented data set (y_i , x_{1i} , x_{2i} , x_{3i} , x_{4i} , z_i)

4. perform best subsets regression on the augmented data set
5. repeat steps (2) to (4) N times and determine the proportion of times that Z is included in subset of size j.

For brevity we merely note that the distribution of the proportion of times Z enters a best subset solution is dependent on the extent of the correlation between the predictor variables. For instance, for a best subset of size j = 1, the distribution of the proportion of the times that Z enters the solution with independent predictors has a mean of 0.2, with simulation estimated quartiles ($Q1$, $Q2$, $Q3$) = (0.06, 0.18, 0.32). However for the correlation structure (2) outlined in Section 4 the corresponding mean was estimated to be 0.26, and the corresponding quartiles were estimated to be ($Q1$, $Q2$, $Q3$) = (0.10, 0.24, 0.38).

9. CONCLUSIONS

A computer based heuristics that allows the Type I error rate for a best subsets regression to be controlled at any pre-determined nominal significance level has been described. The data sets created under the randomization procedure as described precisely retain the correlation structure as observed in the original data and this is critical to the approach.

The outlined procedure corrects a known problem with best subsets regression. The given procedure corrects the bias in the overall p-value for best subsets regression. The correction is a non-trivial correction and even applies in those particularly problematic situations when the number of predictors exceeds the number of cases.

Significance tests in classical least squares regression are based on the assumption that the underpinning error terms are independent identically distributed normal random variables. Even when these assumptions are satisfied the p-values when estimated under best subsets regression are still biased, leading to wrong inferences. This is not the case with the outlined randomization procedure.

In practice the underpinning normality assumptions are likely to be violated to some extent leading to further bias in the p-values in best subsets regression. In contrast the fake variable approach is based on the sample data and the estimation of the p-value does not explicitly rely upon distributional assumptions. The simulation work indicates that the randomization approach retains good statistical properties in the presence of an outlying and/or influential observation. In principle the same procedure could be used for other best subsets regression techniques (e.g. logistic regression models, or more generally the generalized linear) or for other best subsets models (e.g. discriminant analysis).

The fake variable approach, as described, is a randomization approach that preserves the sample correlation structure observed between predictors in each and every fake variable solution. In these respects the newly proposed procedure is different from bootstrapping which is based on random sampling with replacement from sample cases. For a more detailed explanation of bootstrapping techniques see [3].

Stoppiglia *et. al.* [7] and Austin and Tu [1] have considered the use of a single fake variable (also known as a probe variable) to help determine the reliability of any final model. Stoppiglia [7] considers the problem of building a model many times over to determine the ranking of an independent random fake variable in relation to other variables in the model. The rationale is to retain those potential predictor variables that consistently rank higher than the fake variable that "probes" the solution. Austin and Tu [1] do something similar and include a randomly generated single fake variable in each of their bootstrap samples and then determine the proportion of times the fake variable is included in any final bootstrap model for comparison with the proportion of inclusion of the other variables. Note however neither [1] nor [7] give explicit decision rules for the use of a single fake variable. The simulation work outlined in Section 8 suggests that this intuitively appealing approach may be problematic for model building because the frequency of including a probe in solution is problem dependent. This does not necessarily invalidate the use of a probe for validation or use in confirmatory studies. Our work suggests that a more fruitful approach may be to work with multiple fake variables.

More generally the method given in this paper is strongly suggestive of ways in which computer scientists can generate other fake variable algorithms to be used with other heuristics (e.g. backward elimination) and for use with other generic models (e.g. regression trees) and in doing so validly control error rates.

REFERENCES

[1] Austin P. C. and Tu J. V., "Automated Variable Selection Methods for Logistic Regression Produced Unstable Models for Predicting Acute Myocardial Infarction Mortality," *Journal of Clinical Epidemiology*, 57, 1138-1146, 2004.

[2] Derksen S. and Keselman H. J., "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables," *British Journal of Mathematical and Statistical Psychology*, vol. 45, 265-282, 1992.

[3] Efron B. and Tibshirani R. "*An Introduction to the Bootstrap*," Chapman and Hall, 1993.

[4] George E., "The Variable Selection Problem," *Journal of the American Statistical Association*, vol. 95, 1304-1307, 2000.

[5] Miller A. J., "*Subset Selection in Regression*," Chapman Hall, 1990.

[6] Rencher A. C. and Pun F. C., "Inflation of R^2 in Best Subset Regression," *Technometric*s, vol. 22, 49-53, 1980.

[7] Stoppiglia H., Dreyfus G., Dubois R., and Oussar Y., "Ranking a Random Feature for Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, 1399-1414, 2003.