ISSN 1022-8594
JUJS

Vol. 41, No.1, pp.113-124

One-Step and Two-Step Model Buildings: A Comparison

Md. Siddiqur Rahman* **and Md. Abul Kalam Azad**
Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh

Abstract

Building a multiple linear prediction model is very challenging when the data set contains a large number of candidate covariates as well as a fraction of outliers and other contaminations that are difficult to visualize and clean. One-step model building and two-step model building are two different strategies for linear model selection. One-step model building procedure aims to build up a final prediction model in one step by using step-by-step algorithm such as Backward elimination (BE). Two-step model building is a blend of all possible subset regression and step-by-step algorithms. The first step of this procedure, called short-listing, quickly screens out the less important variables to form a "reduced set" for further consideration. The second step, called segmentation, carefully examines different subsets of the variables in the reduced set to build a final prediction model based on only the chosen imported ones. The classical one-step and two-step model building procedures yield poor results when the data contain outliers and other contaminations. Robust version of one-step and two-step model building procedures is introduced in this study. An extensive simulation study is conducted to compare the performance of one-step model building with two-step model building. According to the simulation study and real data application, the two-step model building procedures perform better than the one-step model buildings.

Keywords: Backward elimination, Contamination, Short-listing, Segmentation, Winsorized Correlation.

Introduction

When the number of candidate covariates, d, is small, one can choose a linear prediction model by computing a reasonable criterion (e.g., Mallows C_P, AIC, FPE or cross-validation error) for all possible subsets of the predictors. However, as d increases, the computational burden of this approach (sometimes referred to as all possible subsets regression) increases very quickly. Typically, when a large collection of possible covariates are available, a parsimonious set is required to select from the large collection for the efficient prediction of a response variable. This is

* E-mail of correspondence: rsiddiq11@yahoo.com

one of the reasons why step-by-step model-building algorithms like Backward elimination (BE) are popular [1 - 5]. The BE was expressed in terms of classical correlation [6].

Though one-step model building algorithms construct better model for large data sets, but these algorithms do not guarantee to take all the important covariates in the models. They may also select some noise (wrong) covariates in the models. In two-step model building procedure, the candidate covariates are sequenced by BE. A learning curve is obtained by plotting robust R^2 [7] values to form a list against the number of variables in the model. An appropriate size of the short list can be selected at the point where the learning curve starts to leave off. All subsets selection to the predictors of the short list is applied using the k-fold $100\alpha\%$-trimmed cross validation (CV) procedure on data [8]. The subset that produces the least prediction error is the final model.

Classical BE and CV algorithms give poor results when the data contain a fraction of contamination. Robust BE procedure has been developed which aims to build up a final prediction model in one step using partial F test criteria [6]. The k-fold $100\alpha\%$-trimmed cross validation (CV) procedures was developed to obtain a subset of predictors in the final model that produces the least prediction error [8].

In the following part of the paper, the BE and Robust Backward Elimination (RBE) procedures have been reviewed as one-step model building. The two-step model building procedures has been discussed. The results of a simulation study have been presented and the performance of one-step and two-step model building procedures has been compared. Two real data applications have been demonstrated. The limitations of one-step and two-step model building procedures are also discussed. Finally a conclusion is made.

One-Step Model Building

The BE algorithm: The BE procedure starts with the full model, and removes one covariate at each step. Let r_{jY} denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates. The predictors that are in the current regression model are called “active”

predictors. Suppose without loss of generality X_1 has the minimum absolute partial correlation with Y after eliminating the linear effect of $X_2, X_3, \cdots, X_d$ on X_1. Then, X_1 is the first variable that is dropped from the regression model. This candidate predictor is called "inactive" predictor. Thus to find out the inactive predictor (say, X_1), the partial correlation between X_1 and Y after eliminating the linear effect of $X_2, X_3, \cdots, X_d$ on X_1 is required to compute which is denoted by $r_{1Y.23\cdots d}$. These partial correlations were expressed in terms of original correlations [6]. That is, BE algorithm is formulated in terms of sample means, variances and correlations. Once the correlation matrix is calculated, the actual observations are not required any more.

BE algorithm is summarized in terms of correlations among the original variables as follows.

1. Let D be the set of all covariates and P be the subset not containing jth covariate. To remove the first covariate, X_{m1}, calculate partial correlation $r_{jY.P}$ between X_j and Y after eliminating the linear effect of covariate belonging to P on X_j. Determine $m_1 = \arg\min \left| r_{jY.P} \right|$.

2. Let C be a subset containing $(k-1)$ variables that has been removed from D after $(k-1)$ steps $(k = 2, 3, \cdots)$ and P be the subset not containing jth covariate and C. To remove the kth covariate X_{mk}, $r_{jY.P}$ between X_j and Y may be calculated after eliminating the linear effect of $X_{m1}, X_{m2}, \cdots, X_{m(k-1)}$ on X_j, and then determine $m_k = \arg\min \left| r_{jY.P} \right|$.

At each BE step, once the "weakest" covariate (among the remaining covariates) is identified, we can perform a partial F -test to decide whether to drop this covariate from the model (and continue the process) or to stop. The new "weakest" covariate drops from the model only if the partial F -value, denoted by F_{partial}, is smaller than 90th percentile of F distribution, $F(0.90, 1, n-k-1)$ (say), where k is the current size of the model excluding

the new covariate. Here again, the required quantities can be expressed in terms of correlations among the original variables, as we show below.

When k covariates $X_1, X_2, \cdots, X_k$ are in the model, and without loss of generality X_k has the smallest absolute partial correlation with Y after adjusting X_k for $X_1, X_2, \cdots, X_{k-1}$, the partial F -statistic for X_k can be expressed as

$$F_{\text{Partial}} = \frac{(n-k) r^2_{kY.123\cdots(k-1)}}{1 - r^2_{1Y} - r^2_{2Y.1} - r^2_{3Y.12} - \cdots - r^2_{kY.123\cdots(k-1)}} \tag{1}$$

The RBE algorithm: A simple robustification of BE algorithm (RBE) can be achieved as FS (Forward Selection) and LARS (Least Angle Regression) simply by replacing the non-robust ingredient of this algorithm by their robust counterparts. That is, the initial correlation matrix is computed using adjusted Winsorization method [9] or Spearman's rho that is resistant to bivariate outliers [10]. We call RBE based on adjusted Winsorization correlation as RBEw and RBE based on Spearman's rho as RBEr. The classical correlations are replaced in the partial F statistic by their robust counterparts to form a robust partial F statistic.

Two-Step Model Building

The two-step model building procedure contains two consecutive procedures: short-listing and segmentation.

Short-listing: Sequencing aims to first sequence all the candidate covariates to form a list in which more important ones are likely to appear at the beginning. The first m covariates of the sequenced list will form the short list which will be studied further. A suitable step-by-step algorithm is required to sequence all the candidate covariates. To sequence the candidate covariates, Rahman and Khan proposed an algorithm called BE [6]. The sequence generated by BE is not robust against outliers. So RBE can be used to sequence the candidate covariates [6].

Based on the sequenced generated from RBE, we can first generate the corresponding “short list” which includes the first m top ranking predictors (equal to or slightly larger than the number of predictors in the

final model). But no information is available about the number of predictors needed in the model. A graphical tool may be useful to select the size of the short list. First a robust regression model is fitted taking only the first covariate from the sequence as predictor, and then another covariate is added in the model (one variable at a time) by following the orders of the covariates in the sequence. Each time the number of variables is increased (along the sequence) and a robust regression model is fitted each time to compute a robust R^2 measure such that $R^2 = 1 - \text{med}(e^2)/\text{mad}^2(Y)$, where e is the vector of residuals obtained from the corresponding robust fit [7]. Then the learning curve (recall curve) is obtained by plotting these robust R^2 values against the number of variables in the model. The size of the short list m, can be selected a the point where the learning curve does not have a considerable (increasing) slope anymore.

Segmentation by RCV: Cross-validation (CV) is a method of estimating the error rate of a prediction rule. This estimate is obtained by splitting the n data points into a training sample of size n_t (used for fitting a prediction model, i.e., model construction) and a validation sample of size $n_v = n - n_t$ reserved for assessing the predictive ability of the model. Often k-fold CV is used which means that the data set is split randomly into k blocks of approximately equal size. The training sample then consists of $k-1$ blocks and the validation sample is given by the left-out block. Each block is left out once, so that a prediction is obtained for each of the observations in the sample. The average prediction error is calculated based on a number of possible random k-fold splits of the data set. Suppose that m predictors are selected as a short list. We apply all-subsets selection to these m predictors using k-fold $100\alpha\%$-trimmed cross validation (CV) procedure on data [8]. To robustly measure prediction error, we use regression MM estimator (M estimator is an extension of MLE and is a robust estimator. MM estimator is the development of M estimator). The subset that produces the least prediction error is the final model.

Simulations

An extensive simulation study similar to Rahman and Khan (2014) is carried out to compare the performance of one-step model building and two-step model building procedures [6]. The total number of candidate

covariates is 40, in which $a = 6$ are nonzero covariates. Two different cases according to the different correlation structures among the target covariates: "no correlation" case and "moderate correlation" case are considered which are discussed below.

For the no correlation case (a true correlation of 0 between the covariates), independent predictors $X_j \sim N(0,1)$ are considered, and response variable Y is generated using the a nonzero covariates with coefficients $(7, 6)$ repeated three times for $a = 6$.

For the moderate-correlation case, two latent variables $L_i, i = 1, 2$ are considered to generate $Y = 7L_1 + 6L_2$, where $L_i \sim N(0,1)$, and ε is a normal error not related to the latent variables. The nonzero covariates are divided in three equal groups, with each group related to exactly one of the latent variables by the following relation

$$X_j = L_i + \delta_j$$

where $\delta_j \sim N(0,1)$. Thus, a true correlation between the covariates generated with the same latent variable is 0.5.

For each case, we generated 100 data sets each of which was randomly divided into a training sample of size 100 and a test sample of size 100. Each training data set was then contaminated as follows. To create bivariate outliers, a number of rows (1%) were chosen randomly, and for these rows the covariates values were replaced by large positive numbers, then the corresponding response values were replaced by large numbers.

We used one robust algorithm (RBEw) as one-step and two-step model building procedures on the cleaned and contaminated training data to select and fit the final models and used these models to predict the test data outcomes. In the second step of the two-step model building, five-fold robust cross validation (RCV) procedure is used. For each simulated data set, the 10% trimmed mean of squared prediction error on the test sample was recorded.

Table 1: Performance of the one-step model building and two-step model building procedures in clean and contaminated data for no-correlation and correlation cases

Case	Data	Method	$a = 6$		
			RBEw		
			MSPE	Noise	Target
No-correlation	Cleaned	One-step	49.9 (8.1)	12.0 (0.4)	6.0 (0.1)
		Two-step	31.5 (5.7)	0.0 (0.4)	6.0 (0.1)
	Contaminated	One-step	64.5 (13.9)	8.1 (3.0)	6.0 (0.1)
		Two-step	54.7 (6.0)	3.5 (0.5)	6.0 (0.1)
Correlation	Cleaned	One-step	69.2 (5.5)	6.0 (1.0)	5.5 (0.5)
		Two-step	58.0 (3.9)	2.5 (0.6)	5.5 (0.5)
	Contaminated	One-step	105.7 (30.6)	4.4 (0.5)	5.2 (2.1)
		Two-step	91.9 (15.9)	2.4 (1.3)	4.3 (1.8)

Table 1 shows the average (SD) of mean squared prediction error (MSPE) on the test set, the average number of noise variables (Noise) and the average number of target variables (Target) for one-step and two-step model building procedures by each algorithm. For no-correlation case in clean and contaminated data, the test errors produced by two-step model building procedure are much smaller than the one-step model building procedure. The average (SD) of the third quantity (total number of target variables) is similar for both the methods for clean and contaminated data. Also the models obtained by two-step procedure contain fewer noise variables than the one-step procedure. For correlation case in clean and contaminated data, the test errors produced by two-step model building procedure are much smaller than one-step model building procedure. Two-step model building has taken less noise covariates than one-step procedure. In this case, two-step model building procedure has selected fewer target covariates than one-step model building procedure. For contaminated and correlated data, we have considered first 30% (12) covariates as a short list for segmentation. If we would consider first 35% to 40% covariates as a short list for segmentation, two-step model building could take all the target covariates in its model. If we would consider 40% (16) instead of

30% (12) of the covariates as a short list, the required time would be multiplied by $\frac{2^{16}\times 16^2}{2^{12}\times 12^2}$ or 28.44 to run a simulation.

Applications to real data

In this section, two real-data examples are used to compare the performance of one-step and two-step model building procedures.

Brain-Computer Interface (BCI) Data: This data set was used in the BCI competition III (data set V). It represents time series of *electroencephalogram (EEG) signal readings. We consider train_subject01_psd04 data set which consists of* $n = 3488$ data points. There are 97 variables in the data: The first 96 are continuous variables and the last one a numerical level. We consider the first 31 variables in our analysis, and use the second variable as the response. Thus, there are 30 covariates and one response in the selected data. RBEw (with $F_{0.90}$ as the deletion criterion) applied to this data set (including outliers) selected the following model of 15 covariates:

2 1 3 4 29 12 26 23 6 11 20 19 17 14 21

We fitted the selected model using the training data, and then used them to predict the test data outcomes. The 5% (10%) trimmed MSPE is 0.000042(0.000033).

Again, RBEw applied to this data set (including outliers) produced the sequence (2 1 3 4 29 12 26 23 6 11 20 19 17 14 21 5 24 9 15 10 18 13 27 7 8 25 30 16 22 28). We used this sequence and fitted Least Median Squares regression to obtained robust R^2 values [11]. Figure 1 shows the learning curve for the BCI data based on the above sequence.

This plot suggests a short list which includes the covariates (2 1 3 4 29 12 26). The short lists will be used in the segmentation step to build up the final model. We applied robust segmentation five-fold (10% trimmed) CV (RCV) method using MM-estimator on the above short list. The covariates selected in the final model are (2, 1, 3, 4). We fitted this model using the training data, and then used them to predict the test data outcomes. The 5% (10%) trimmed MSPE for the model is 0.000041 (0.000033) which is same

as one-step model. It is clear that the two-step model building selects vary fewer covariates in its model compared to its corresponding one-step model building.

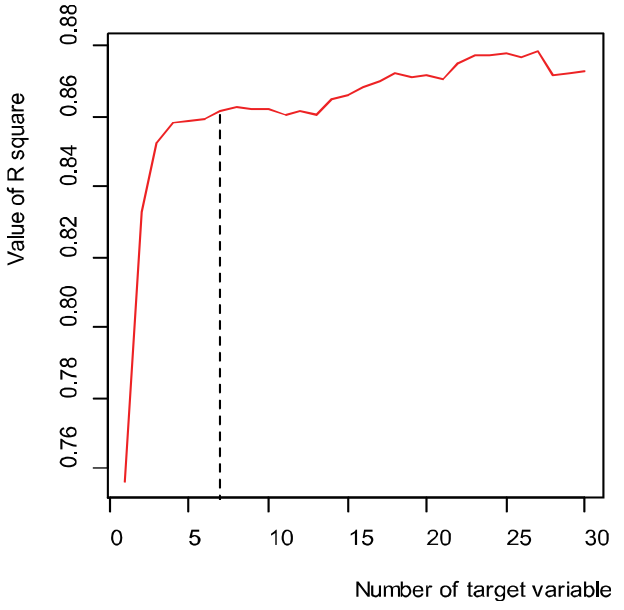

Figure 1: Learning curve for BCI data

Protein data: This data set of $n = 145751$ protein sequences was used for KDD-Cup 2004. Each of the 5 blocks corresponds to a native protein, and each data-point of a particular block is a candidate homologous protein. We considered first $n = 4565$ protein sequences from 5 blocks. The number of covariates is $d = 77$. The first 2 are indicator variables (variables 1-2), the third is the class (Proteins that are homologous to the native sequence are denoted by 1, non-homologous proteins by 0), and rest 74 are the features variables. The first feature is considered as a target variable. Thus, there are 73 covariates. The data were split to get a training sample of size $n = 2280$ and a test sample of size $n = 2285$. RBEr (with $F_{0.90}$ as the deletion criterion) applied to this data set (including outliers) selected the following model of 39 covariates:

(9, 1, 11, 65, 14, 4, 10, 66, 42, 32, 21, 20, 72, 62, 73, 49, 27, 29, 17, 18, 25, 50, 51, 43, 47, 48, 33, 61, 60, 7, 16, 63, 68, 71, 3, 2, 69, 67, 57)

We fitted the selected model using the training data, and then used them to predict the test data outcomes. The 5% (10%) trimmed MSPE is 100.1 (86.5).

Again, RBEr applied to this data set (including outliers) and the resulting learning curve (for first 30 variable of the sequence) is shown in Figure 2.

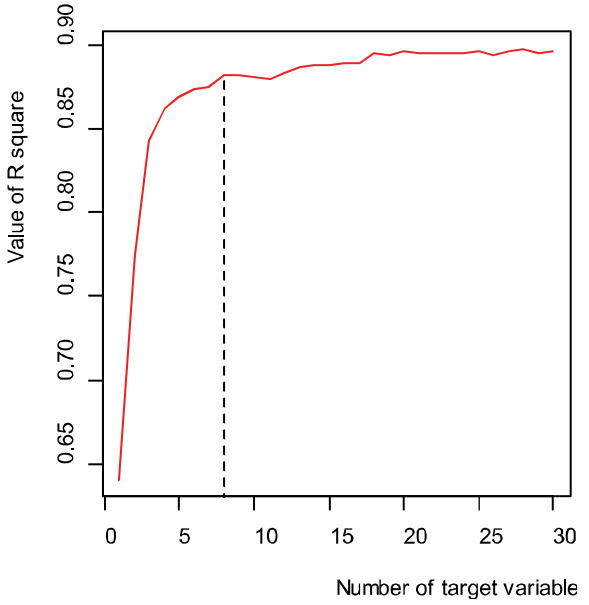

Figure 2: Learning curve for protein data

This plot suggests a very short list of at most size 8 which includes the covariates

(9, 1, 11, 65, 14, 4, 10, 66).

Using five-fold (10% trimmed) RCV procedure yields the final model with seven covariates (1, 4, 9, 11, 14, 65, 66). The 5% (10%) trimmed MSPE for this model is 101.2 (88.4), which is almost same as MSPE for one-step model.

Limitations

RBE is resistant to bivariate (correlation) outliers. However, it may be sensitive to three or higher-dimensional outliers, that is, outliers that are not

detected by univariate and bivariate analyses. Also, the correlation matrix obtained from adjusted Wisorized correlation or Spearman's rank correlation approach may not be positive definite, forcing the use of correction for positive definiteness in some cases [12]. Though one-step model building algorithms construct better models for large data sets, but these algorithms don't guarantee to take all the important covariates in the models. They may also select some noise (wrong) covariates in the models. From the simulation studies, it is evident that the performance of two-step model building procedure is better than one-step model building procedure. However, in several occasions, even the proposed two-step procedure selects some noise variables in the model in addition to correct variables.

Conclusion

BE is a popular and computationally suitable algorithm for building linear prediction models. BE has been expressed in terms of Pearson's product moment correlations [6]. The BE is very sensitive when the data contain contaminations (gross errors or deviations from normality). Since adjusted Winsorized correlation and Spearman's ρ are more reliable estimates of association in the presence of contaminations in the data, they have been introduced in BE. That is, replacing Pearson's product moment correlations by adjusted Winsorized correlations and Spearman's ρ in BE algorithm, the respective BEw and BEr have been obtained. RBE (BEw or BEr) procedure aims to build up a final prediction model as one-step model building algorithm using partial F test criteria. Again, in two-step model building, all the covariates have been sequenced by RBE algorithm and a short-list has been obtained from learning curve. In the segmentation step, computationally suitable robust version of CV (RCV) (evaluation of all subset of the short list) has been applied to select the final model. According to simulation study and real data application, it is evident that the two-step model building procedure performs better than the one-step model building procedure.

References

[1] G. Furnival and R. Wilson, *Regression by Leaps and Bounds*, Technometrics, 1974, vol. 16, pp. 499-511.

[2] C. Gatu and E.J. Kontoghiorghes, *Branch-and-bound algorithms for computing the best subset regression models*, Journal of Computational and Graphical Statistics, 2006, vol. 15, pp. 139-156.

[3] Huo and Ni, *When Do Stepwise Algorithms Meet Subset Selection Criteria? The Annals of Statistics*, 2007, vol. 35, pp. 807-887.

[4] Das and Kempe, *Algorithms for subset selection in linear regression*, Proceedings of the 40th annual ACM symposium on Theory of computing, ACM, New York, NY, USA, 2008.

[5] S. Weisberg, *Applied Linear Regression (2nd ed.)*, Wiley, New York, 1985.

[6] M. S. Rahman and J. A. Khan, *Building a Robust Linear Model with Backward Elimination Procedure*, The Dhaka University Journal of Science, 2014, vol. 62, pp. 87-93.

[7] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley-Interscience, New York. 1987.

[8] J. A. Khan, S. Van Aelst and R. H. Zamar, *First robust estimation of prediction error based on resampling*, Comp. Statist. Data Anal, 2010, vol. 52, pp 239-248.

[9] J. A. Khan, S. Van Aelst and R. H. Zamar, *Robust Linear Model Selection Based on Least Angle Regression*, J. Amer. Statist. Assoc, 2007(b), vol. 102: pp.1289-1299.

[10] M. S. Rahman, *Backward Elimination Procedure for Linear Model Building Using Spearman's Rank Correlation*' Jahangirnagar University Journal of Science, 2015, vol.38, pp. 11-22.

[11] P. J. Rousseeuw, *Least Median of Squares Regression*, Journal of the American Statistical Association, 1984, vol. 79: pp. 871–880.

[12] F. A. Alqalla, K. P. Konis, R. D. Martin, and R. H. Zamar, *Scalable Robust Covariance and Correlation Estimates for Data Mining*, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, 2002, pp. 14-23.