The International Journal of Biostatistics

Volume 5, Issue 1

2009

Article 25

On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression

Zafar Mahmood, *NWFP Agricultural University, Peshawar*
Salahuddin Khan, *University of Peshawar*

Recommended Citation:

Mahmood, Zafar and Khan, Salahuddin (2009) "On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression," *The International Journal of Biostatistics*: Vol. 5: Iss. 1, Article 25.

DOI: 10.2202/1557-4679.1105

On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression

Zafar Mahmood and Salahuddin Khan

Abstract

This paper addresses a methodological technique of leave-many-out cross-validation for choosing cutoff values in stepwise regression methods for simplifying the final regression model. A practical approach to choose cutoff values through cross-validation is to compute the minimum Predicted Residual Sum of Squares (PRESS). A leave-one-out cross-validation may overestimate the predictive model capabilities, for example see Shao (1993) and So et al (2000). Shao proves with asymptotic results and simulation that the model with the minimum value for the leave-one-out cross validation estimate of predictor errors is often over specified. That is, too many insignificant variables are contained in set β_i of the regression model. He recommended using a method that leaves out a subset of observations, called K-fold cross-validation. Leave-many-out procedures can be more adequate in order to obtain significant and optimal results. We describe various investigations for the assessment of performance of predictive regression models, including different values of K in K-fold cross-validation and selecting the best possible cutoff-values for automated model selection methods. We propose a resampling procedure by introducing alternative estimates of boosted cross-validated PRESS values for deciding the number of observations (l) to be omitted and number of folds/subsets (K) subsequently in K-fold cross-validation. Salahuddin and Hawkes (1991) used leave-one-out cross-validation to select equal cutoff values in stepwise regression which minimizes PRESS. We concentrate on applying K-fold cross-validation to choose unequal cutoff values that is F-to-enter and F-to-remove values which are then used for determining predictor variables in a regression model from the full data set. Our computer program for K-fold cross-validation can be efficiently used for choosing both equal and unequal cutoff values for automated model selection methods. Some previously analyzed data and Monte Carlo simulation are used to evaluate the proposed method against alternatives through a design experiment approach.

KEYWORDS: cross-validation, cutoff values, stepwise regression, prediction, variable selection

Author Notes: The authors wish to thank the referees for their constructive comments, especially the editor's careful reading of the paper and helpful suggestion, which led to several substantial modifications in the presentations.

1. Introduction

1.1. Assessment of variable selection in regression models

In many research areas such as biological, medical, public health, social, and agricultural sciences, there may be large number of predictor variables that could be used to predict a regression model. If the number of predictor variables are denoted by p , various variable selection procedures are often define by selecting $r < p$ predictor variables that allow to construct a best subset regression model. We select the subset regression model with “ r ” predictor variables due to the following major reasons.

- I) It is very simple and cheap to select r predictor variables in the final regression model.
- II) Accuracy of regression model is increased due to the exclusion of unnecessary predictor variables.
- III) Calculations of estimated regression coefficients are potentially fast when less number of input variables is used in the regression model.
- IV) The nature of the prediction problem can be recognized by knowing only relevant predictor variables in the regression model.

The number of subset model grows exponentially with the number of predictors. Thus it is difficult to evaluate all the subset models even for moderate number of predictor variables. Many heuristic algorithms have been proposed for determining subset models; for example, see Stearns (1976), Kittler (1978), Pudil et al. (1994), Somol et al. (1999) and Somol & Pudil (2000).

In variable selection, the main objective is to choose the subset model having small number of predictor variables to enable the construction of accurate predictive regression model. Consequently, the accuracies of the selected predictors are validated to check that a good subset model has been finalized. Model validation ascertains whether predictor values from the model will predict accurate responses on future subject or subjects not used in the model. Usually one select predictor variables using subset of the data samples available for training (called training set) and test them with the rest of the samples (called test set). A common choice is cross-validation, where the data sample is randomly divided into a number of folds.

1.2. Prediction models and its validation

A regression model is not useful unless one has an estimate of the predictive ability. We consider the classical problem of the multiple regression models to check the behavior of a dependent variable Y by a linear function of predictor variables $X_1, X_2, \dots, X_p$ that is,

$$Y = X\beta + \varepsilon$$

Where Y is an $n \times 1$ vector of response variable observations, X is an $n \times p$ known matrix of p explanatory variables, β is a $p \times 1$ unknown coefficients but constant vector, and ε is an $n \times 1$ random error vector assumed to be distributed $N(0, \delta^2_{1n \times n})$. Here we have a random sample of $(p+1)$ -tuples $\{x_{i1}, x_{i2}, \dots, x_{ip}, y_i\}$, where $i = 1, 2, \dots, n$ or $\{x_i, y_i\}$, where $i = 1, 2, \dots, n$ as the training set.

Multiple regression analysis uses the ordinary least square estimates for β given by $\hat{\beta} = (X'X)^{-1}X'Y$ and thus estimates the regression model. Also the unbiased estimate of the variance δ^2 is given by $S^2 = e'e / (n-p)$, where e is the vector of residuals $(Y - X\hat{\beta})$. The basic need is to assess the performance of this prediction and to validate it. One way is to test the model on set of m further observations from the same population, $\{x_i, y_i\}$, where $i = n+1, n+2, \dots, m$ say and then compare the observed values of Y_i with those of predicted values $\hat{Y}_i$ obtained from the regression model. This can be done by calculating standard

prediction sum of squares as criterion, Standard PRESS =
$$\frac{\sum_{i=n+1}^{n+m} (Y_i - \hat{Y}_i)^2}{\sum_{i=n+1}^{n+m} (Y_i - \bar{Y})^2}$$
, Where

$\bar{Y}$ is the mean of the dependent variable in the test set. But if test set is not available, the model can be validated by comparing the prediction $\hat{Y}_i, i = 1, 2, \dots, n$ of the training set with their counterparts Y_i . However, this will give biased values if matching does not occur for independently collected data. Thus a favored alternative is cross-validation in which the procedure is to split the available data into equal parts (see Luchenbruch and Mickey (1968) and Stone (1974)). One part of the data will be used for estimating the regression coefficients called the estimation data or training data while the remaining part will be used for prediction purpose of the model called prediction data or test data.

1.3. Cross-Validation, leave-one-out and K-fold classification

Cross-validation is a computer intensive resampling technique of applied statistics used for the assessment of statistical models and selection amongst competing regression models. Basically, it is a standard tool for estimating prediction errors of statistical models. Snee (1977) claimed that cross-validation is an effective method of model validation when it is not possible for the researcher to collect new data. A considerable work has been written on both the theoretical and practical aspects of cross-validation, for example, see Stone (1974), Geisser

(1975), Shao (1993), Breiman (1995) etc. The idea of cross-validation is simply splitting the data set into two parts, using one part called the training set to derive a prediction rule and then judge the goodness of the prediction by matching its out with rest of the data called test set. There are several versions of cross-validation (CV), e.g. leave-one-out cross-validation (LOOCV, also called simple CV), leave-many-out cross-validation (LMOCV, also called K-fold CV) and generalized CV (Cravan and Wahba (1979)). However, in the literature unless indicated explicitly cross-validation is usually referred to as the leave-one-out cross-validation.

As practically cross-validation (CV) is computationally very expansive, so we search for less greedy CV algorithm such as K-fold cross-validation (KFCV). Efron (1986) argued that simple cross-validation (that is LOOCV) is a poor candidate for estimating the prediction error and suggests that some version of bootstrap as better alternative. When selecting the correct model is concern, it is well known that the model selected by cross-validation criterion is apt to over fit. Burman (1990) showed that the multifold cross-validation is asymptotically optimal. Herzberd and Tsukanov (1986) have provided simulations evidence that multiple cross-validations do better than simple cross-validation.

The properties of leave-many-out cross-validation for prediction and model identification have been widely studied from the asymptotic viewpoint in regression analysis. It basically depends upon the ratio between the size of test and training sets, called the split ratio $l / (l - p)$ in the leave- l -out case and $1/1-K$ for K-fold. van der Laan, Dudoit and Keles (2004) studies asymptotic optimality of a general class of likelihood based cross-validation techniques. They concluded that the cross-validation sector performs asymptotically as good as optimal model selector which depends on true density. Dudoit and van der Laan (2005) presented a general framework for cross-validation and derive distributional properties of cross-validated risk estimators in the context of model selection and performance assistance. Van der Vaart, et al. (2006) considered choosing an estimator or model from a given class by cross-validation and presented finite sample oracle inequalities.

The idea of leave-many-out cross-validation first appear in Geisser (1975) where instead of deleting one observation as in simple cross-validation, $l > 1$ observations are deleted. Some development in this area can be found in Breiman et al. (1984), Bruman (1989). A computationally cheaper version of leave-one-out cross-validation is called K-fold cross-validation, sometime also known as rotation estimation. In this method, the data is partitioned into K-mutually exclusive subsets (the folds) $n_1, n_2, \dots, n_K$ of approximately equal size. Mathematically,

$$n_l \text{ of } \{j=1, 2, \dots, n\} \text{ such that } \bigcup_{l=1}^K n_l = \{1, 2, \dots, n\} \text{ and } n_j \cap n_l = \emptyset \ (j \neq l).$$

Each subset group is omitted in turn from the data and is used as the test set. The remaining $K-1$ subsets are put together and are used as the training set. The regression model is fitted to the training set and the predictions $\hat{Y}_i$ are derived from the test set. Every data point gets to be in a test set exactly once and gets to be in training set $K-1$ times.

The K -fold cross-validation produces n predictions $\hat{Y}_i$, none of which has used the corresponding Y_i as part of the modeling. So the PRESS calculated should not be optimistically biased. The special case where the number of folds is set equal to the number of samples available is again called the leave-one-out cross-validation.

When it comes to the practical application, the main problem is the choice of K (in K -fold CV) or l (in leave- l -out CV) since the performance of CV strongly depends on them. Basically, there are three competing factors (bias, variance and computational cost) to take into account in order to choose these tuning parameters. One of the purposes of this article is to suggest a procedure for selecting appropriate tuning parameter (K or l) for CV procedures.

2. Choice of l (number of individuals in each subset) and K (number of folds) in K -fold cross-validation

K -fold cross-validation is extremely useful if correct value of l that is number of individuals in each fold or appropriate K that is number of folds is chosen. It is computationally less expensive than the leave-one-out cross-validation and also provides the best estimated cross-validation error if the correct value of K is used.

Unfortunately, there is no theoretically perfect procedure of determining the appropriate value of K and number of observations (l) in each K th group. Many researchers used $K = 10$, for example, see Bengio and Grandvalet (2004), Vehtari and Lampinen (2004). Davison and Hinkley (1997) recommend $K = \min(n^{1/2}, 10)$ in practice. Clark (2003) compare different values of K for his data and suggested that the choice of $K = 4$ will probably be good in general, though not necessarily optimal. Duan et al. (2003) used $K = 5$ for the number of folds. Jonathan et al. (2000) recommended the use of leave-2-out cross-validation for all sample size in the interval $[20, p]$ simulation. They also mentioned that for $K > 4$, the analysis of number of different partitioned of the design data set should always be considered. Anderson et al (2006) suggested, leaving 20% of the n -samples at a time for final model validation. Baumann and Stiefl (2004) used leave-50%-out cross-validation. Shao (1993) mentioned that in leave-many-out cross-validation a large portion (40-60%) of the training data is set aside as validation data set. Using the value $K=10$ seems to be a good rule of thumb but the true best value may differ for each algorithm and each data set. Several

researchers have suggested $l = \frac{n}{K}$ where K is number of subsets (folds), n is total number of observations and l is number of observations in each omitted subset, for example see, Jounathan, et al. (2000), Kohavi (1995), Bengio and Grandvalet (2004), Aldrin (2006) etc. Some researchers used $K = [rn]$ with $r = 0.1$ (10% test data).

We propose a bootstrap cross-validated procedure for deciding the number of observations (l) to be omitted and the number of folds/subsets (K) subsequently in K-fold cross-validation for each data set. The procedure can be formulated using the following four steps.

1. Read input data as n -rows (number of observation) and m -columns (number of predictor variables + a response variable).
2. Choose group sizes $n_1, n_2, \dots, n_K$ such that $n_1 + n_2 + \dots + n_K = n$ and initialize the bootstrap cross-validation procedure for any value of $l = 1, 2, 3, \dots$ (We choose for example l as all multiple integers of n).
 - (i) Delete ' l ' number of observations of group n_1 from the data set for each bootstrap sample and fit the full regression model for all ' b ' reduced samples.
 - (ii) Predict the deleted case(s) and compute the square of predicted residual.
 - (iii) Return to step 2(ii) and repeat the procedure by deleting instead group size n_2 , then group size n_3 and so on until all groups have been deleted once.
3. Sum all the partial square of predicted values for each bootstrap sample and Compute the average of the boosted PRESS.

$$PRESS_{boot} = \frac{\sum_{b=1}^B \left[\sum_{i=1}^n (Y_i - \hat{Y}_{(i)})^2 \right]}{B}$$

Where, B is the number of bootstrap samples.

4. Select the ' l ' value that produce a near minimum $PRESS_{boot}$ and find K as $K = \frac{n}{l}$.

Similar to Wisnowski et al. (2003), we proposed relaxing the requirement of minimum $PRESS_{boot}$ with a low (not necessarily minimum) $PRESS_{boot}$. Wisnowski suggested to select a model that has the near optimum prediction error with the fewest number of predictor variables. Our strategy is to select the K -value that has the near minimum $PRESS_{boot}$. This can also be observed by

plotting a scree-type plot (line-connected scatter plot of the $PRESS_{boot}$ values). The smaller K-value in the range of K-values where the screeplot levels off is selected. A practitioner would more likely to adopt our strategy because we chose the temperate K which is neither too large (computational issue arises) nor too small (bias is increased and leads to underfitting).

A scheme of how the bootstrap cross-validation select K in K-fold cross-validation is given in Table 1. We generated artificial data sets of sample sizes $n = 50, 100$ with ten predictor variables and 0, 1, 3, 5, 7 and 10 true predictor variables. The boosted K-fold cross validated prediction sum of squares was recorded using 100 bootstrap samples for each data set.

Table 1: Boosted prediction sum of squares from 100 bootstrap samples for various K values in K- fold cross-validation. The shaded values represent the near minimum boosted prediction error.

Sample Size	l (%)	K	No. of True Predictor Variables.					
			0	1	3	5	7	10
50	50	2	72.905	70.770	92.462	81.576	78.893	65.182
	20	5	54.117	53.930	66.400	60.694	58.384	49.009
	10	10	50.796	51.009	63.015	58.161	55.084	46.387
	4	25	50.274	50.272	62.242	57.673	54.469	46.183
100	50	2	151.8024	130.2620	126.9075	144.5291	129.539	113.2368
	25	4	140.9183	116.6403	117.4841	135.2560	119.5416	104.1548
	20	5	138.1800	114.2909	115.3909	133.8126	116.2503	101.9245
	10	10	135.6080	111.8259	112.2521	129.8925	113.8864	100.2488
	5	20	134.4755	111.5042	112.0684	129.6159	113.6878	98.7117
	4	25	134.1953	111.3538	111.9448	129.6451	113.6027	98.9697
	2	50	134.0572	111.2860	111.8078	129.4134	113.3148	98.5063

Our procedure looks at the change in the boosted prediction sum of squares for various K-values and we select that K-value which is associated with a significantly diminish change in the boosted cross validated prediction sum of squares that is $PRESS_{boot}$. Thus for a good optimal choice we believe that a suitable choice of K is obtainable by relaxing the strict minimum boosted PRESS. Rather, one could choose the K-value for KFCV that has near minimum boosted PRESS. If the change in boosted cross-validated PRESS does not exceed a certain percentage of the boosted PRESS for different K-values, then the larger K-value is selected. The limitations of this procedure is that it work best for relatively low dimensions and we suggest to search best K initially for all the multiple integers of sample size “n” with $p < 15$.

Figure1: Representative Screeplot of boosted CV predicted sum of squares for sample of size n=50

Figure2: Representative Screeplot of boosted CV predicted sum of squares for sample of size n=100

3. K-fold cross-validation as a method for optimizing F_{in} and F_{out} in stepwise regression

K-fold cross-validation is resampling technique and can be used for the choice of subset regression model and for the empirical testing of the reliability of multiple regression models. The aim is achieved by developing a predictive regression model for the training set of the collected data and then using the model to predict the outcome for testing set. In K-fold cross-validation, the data set is divided into K-mutually exclusive subset (the folds). Each subset group is omitted in turn from the data set and is used as the test set. The remaining K-1 subsets are together called as the training set.

Salahuddin and Hawkes (1991) used leave-one-out cross-validation (LOOCV) scheme for choosing equal cutoff values that is $F_{in} = F_{out} = F_0$ in stepwise regression. We have used K-fold cross-validation technique for optimizing both equal and unequal cutoff values that is F-to-enter and F-to-remove values in stepwise regression.

The procedure is that for each possible pair of cutoff values, the predicted values $\hat{Y}_i$ for the test set are calculated from a model selected by stepwise regression method with the subset group is deleted in turn. The predicted residual for each case is calculated and summed to compute the predicted residual sum of squares (PRESS). Finally, those cutoff values are selected which produces the last occurrence minimum value of PRESS.

The idea behind the last occurrence of the minimum PRESS is that we are interested in selecting those cutoff values which produces the minimum PRESS but include the least number of predictors (assumed to be adequate for describing a response variable) in the final regression equation. Because the last value of the minimum PRESS will occur with the larger value of F_{in} , the cutoff values for stepwise regression will be larger and the number of predictor variables in the final model from the full data sets will be smaller.

Algorithm

1. Read input data for n number of data points and $m = p + 1$ number of columns for predictor variables and a response variable.
2. Initiate the cutoff values $F_{in} = 2$ and $F_{out} = 2$ (if equal cutoff values) or $F_{in} = 3$ and $F_{out} = 2$ (if unequal cutoff values).
3. Choose group sizes $n_1, n_2, \dots, n_K$ such that $n_1 + n_2 + \dots + n_K = n$.

4. Initialize the cross-validation procedure for $l = 1, 2, 3, \dots$
 - (a) Delete l observations of group n_1 from the data set and run stepwise procedure using the cutoff values in step 2.
 - (b) Predict the deleted case(s) by using the selected model and compute the square of predicted residuals.
 - (c) Return to (a) and repeat the procedure by deleting instead group n_2 , then group n_3 and so on until all groups have been deleted once.
 - (d) Compute the PRESS values by summing all the partial square of predicted residuals.
5. Increase the cutoff values by increment $dF = 1$ and return to step 4. Repeat the whole procedure and compute the PRESS values. Similarly compute PRESS values for other cutoff values.
6. Select those F_{in} and F_{out} values that produce minimum PRESS. In practice, the minimum PRESS value appears for a range of cutoff values. Applying parsimony principle; we select those cutoff values which produce the last occurrence of the minimum PRESS. As the last occurrence of minimum PRESS occurs with the larger cutoff values and thus smaller number of predictor variables will be selected in the final regression model for the full data set
7. The procedure continue until four increasing steps are completed after the minimum PRESS or F_{in} and F_{out} takes such values that fit the null model.
8. Increase the selected F_{in} and F_{out} values by increment $dF = 0.1$, return to step 4, and repeat the procedure. The cross-validation stops when cutoff values reach a value greater than $F_{in}+0.9$ or $F_{out}+0.9$. Select those cutoff values which produce the last occurrence of the minimum PRESS.
9. Fit the final regression model for full data set by stepwise regression procedure using the selected cutoff values.

4. Case studies

Example 1

To demonstrate the proposed method we used the data given by Kutner et al. (2004). This data is regarding the study of relation of amount of body fat (Y) based on sample of 20 healthy females 25-34 years old. The predictor variables are triceps skinfold thickness (X1), thigh circumference (X2), and midarm circumference (X3). Table 2 lists the PRESS values corresponding to various unequal cutoff values and the plot of the values of PRESS, F_{in} and F_{out} are shown in Figure 3. The final model results are summarized in Table 3.

Table 2: The cutoff values and their corresponding PRESS values

Fent	Frem	PRESS	Fent	Frem	PRESS
3.0	2.0	99.0905	32.0	31.0	99.0905
4.0	3.0	99.0905	33.0	32.0	99.0905
5.0	4.0	99.0905	34.0	33.0	99.0905
6.0	5.0	99.0905	35.0	34.0	99.0905
7.0	6.0	99.0905	36.0	35.0	99.0905
8.0	7.0	99.0905	37.0	36.0	99.0905
9.0	8.0	99.0905	38.0	37.0	99.0905
10.0	9.0	99.0905	39.0	38.0	99.0905
11.0	10.0	99.0905	40.0	39.0	99.0905
12.0	11.0	99.0905	41.0	40.0	99.0905
13.0	12.0	99.0905	42.0	41.0	99.0905
14.0	13.0	99.0905	43.0	42.0	99.0905
15.0	14.0	99.0905	44.0	43.0	99.0905
16.0	15.0	99.0905	44.1	43.1	99.0905
17.0	16.0	99.0905	44.2	43.2	99.0905
18.0	17.0	99.0905	44.3	43.3	99.0905
19.0	18.0	99.0905	44.4	43.4	99.0905
20.0	19.0	99.0905	44.5	43.5	99.0905
21.0	20.0	99.0905	44.6	43.6	99.0905
22.0	21.0	99.0905	44.7	43.7	110.3214
23.0	22.0	99.0905	44.8	43.8	110.3214
24.0	23.0	99.0905	44.9	43.9	110.3214
25.0	24.0	99.0905	45.0	44.0	110.3214
26.0	25.0	99.0905	46.0	45.0	162.2864
27.0	26.0	99.0905	47.0	46.0	162.8962
28.0	27.0	99.0905	48.0	47.0	162.8962
29.0	28.0	99.0905	49.0	48.0	162.8962
30.0	29.0	99.0905	50.0	49.0	162.8962
31.0	30.0	99.0905	51.0	50.0	192.1448

Figure 3: Three dimension plot for Fent, Frem, and PRESS values

Table 3: summary statistics for the final model of body fat data

=====			
Number of variables= 4			
Number of observations=20			
F level to enter= 44.600			
F level to remove= 43.600			
=====			
Constant term for the regression equation= -23.6345			
Variable	Regression Coeff.	STD Error of Coeff.	Partial F-test
-----	-----	-----	-----
X2	0.85655	0.110016	60.617
=====			

We are using this data to illustrate our proposed procedure. Our program selects the unequal cutoff values that is $F_{in} = 44.6$ and $F_{out} = 43.6$ corresponding to the last occurrence of the minimum $PRESS = 99.0905$ and thereafter using these cutoff values to select the final regression model. The procedure can also be run to choose equal cutoff value (our procedure giving $F_0 = 44.6$) for both entry and removal of predictor variables corresponding to the minimum $PRESS$. But some people prefer to make F-to-remove slightly smaller than F-to-enter to introduce a

small bias to keep a variable in the model once it has been entered, for example, see Glantz and Slinker (2000). We suggest selecting the final model on the basis of unequal optimized cutoff values corresponding to the last occurrence of minimum PRESS. The idea behind the last occurrence of the minimum PRESS for a range of cutoff values is that we are interested in selecting those cutoff values that have minimum prediction sum of squares but include the least number of predictor variables (assumed to be adequate for describing the output variable) in the final regression model. We used the last occurrence of the minimum PRESS because this will occur with the larger cutoff value for entry and the number of predictor variables in the final model obtained from full data set will be smaller.

Example 2

The data for this example was given by Kutner et al. (2004). This data set is concerned with the study of predicting survival time in patients undergoing a particular type of linear operation. 54 patients are selected with 8 predictor variables; blood clotting score (X1), prognostic index (X2), enzyme function test score (X3), lever function test score (X4), age in years (X5), gender with 0 = male and 1 = female (X6), and two design variable for the history of alcohol used, (X7,X8). To make the distribution more nearly normal, logarithmic transformation is used by calculated $\log Y$ for the survival time response variable. Table 4 lists the PRESS values corresponding to various unequal cutoff values and the plot of the values of PRESS, F_{in} and F_{out} are shown in Figure 4. The final model results are summarized in Table 5.

Table 4: The cutoff values and their corresponding PRESS values

Fent	Frem	PRESS	Fent	Frem	PRESS
3.0	2.0	2.3178	10.4	9.4	2.4588
4.0	3.0	2.2653	10.5	9.5	2.4505
5.0	4.0	2.3586	10.6	9.6	2.4505
6.0	5.0	2.3619	10.7	9.7	2.4631
7.0	6.0	2.3272	10.8	9.8	2.4531
8.0	7.0	2.3272	10.9	9.9	3.5064
9.0	8.0	2.3272	11.0	10.0	3.5064
10.0	9.0	2.3272	12.0	11.0	3.6885
10.1	9.1	2.2403	13.0	12.0	3.7702
10.2	9.2	2.4588	14.0	13.0	5.1757
10.3	9.3	2.4588			

Figure 4: Three dimension plot for Fent, Frem, and PRESS values for surgical unit data

Table 5: summary statistics for the final model of surgical unit data

```
=====
Number of variables= 9
Number of observations=54
F level to enter= 10.100
F level to remove= 9.100
=====
Constant term for the reg. equation=      3.8524
```

Variable	Regression Coeff.	STD Error Of Coeff.	Partial F-test
-----	-----	-----	-----
X 3	.01545	.001396	122.597
X 2	.01419	.001731	67.182
X 8	.35297	.077191	20.909
X 1	.07332	.018973	14.935
=====	=====	=====	=====

5. Simulation studies

We conducted simulation studies to evaluate the performance of our procedure systematically. Altogether, three simulation studies were conducted and each consisting of 100 runs. In each simulation per run, we generate;

- (1) A total of 550 standardized random observations were randomly generated in a data matrix consisting of $n=50$ rows and $m=11$ columns. The 1st 10 columns were considered as predictor variables and 11th column was taken as response variable.
- (2) A matrix of $n=50$ rows and $m=30$ columns were created from 1500 independently generated standardized normal observations. The 1st 29 columns were taken as predictor variables and the 30th column was taken as a response variable.
- (3) A matrix was created with $n=20$ rows and $m=20$ columns from independently randomly generated standardized normal observations. As usual the 1st 19 columns were taken as predictor variables and the 20th column as a response variable.

The above three simulation studies (where the predictors and the response are assumed to be independent that is all predictors are non-predictive variables) were further investigated with data sets having some true regression parameters. Similar to Gunst and Mason (1980) data, we generated data sets with some known significant predictor variables. We examined our procedure on data sets with true predictor variables, $n\text{TrueVar's} = 1, 2 \text{ and } 3$ respectively while the effect of the variables in the true model was chosen as to be rather small that is $\text{effectSize} = 1$.

We compared four variable selection algorithms: stepwise AIC, Stepwise BIC, classical Stepwise with the default cutoff values equals 4 and stepwise method with the cutoff values we have optimized applying K-fold cross validation.

For stepwise AIC, we used the function `stepAIC` in R (R development core team (2004))- By replacing the default penalty size 2 with $\log(n)$, the function `stepAIC` can be used to perform stepwise search using the BIC rather than AIC as the criterion. For our cross-validated stepwise we made a code in FORTRAN (91) which can be efficiently used for optimizing cutoff values and then selecting the final regression model. Table 6 report the success rate (based on 100 trials) that each method is able to find the correct subset model. We see that our method select the true model for all these three different simulation studies with a sufficiently high probability in comparison with other methods that is classical stepwise method, AIC and BIC based methods.

Table 6: Simulation study: Success rate based on 100 repetitions

Method	Simulation											
	I				II				III			
	No. of True Variables				No. of True Variables				No. of True Variables			
	0	1	2	3	0	1	2	3	0	1	2	3
CSW	0.62	0.72	0.40	0.60	0.38	0.20	0.30	0.20	0.38	0.30	0.30	0.30
StepAIC	0.20	0.30	0.10	0.39	0.15	0.00	0.00	0.00	0.02	0.00	0.00	0.00
StepSBC	0.56	0.63	0.54	0.61	0.30	0.12	0.23	0.31	0.02	0.00	0.00	0.00
OurMethod	0.90	0.96	0.92	0.93	0.93	1.00	0.90	0.90	0.72	0.60	0.50	0.45

6. Conclusion

Cross-validation can play a useful practical role in model selection among the competing regression models and choosing the appropriate cutoff values in stepwise regression. The conceptual framework is simpler than the typical diagnostic procedures as the models are directly judged on their predictive ability. The simplicity of framework makes it directly applicable to a wide variety of practical problems. The development of new and the evolution of known learning algorithms increases the need for estimation methods and fitting regression models which can produce more reliable estimates.

In this study, K-fold cross-validation is applied to optimize cutoff values that are F-to-enter and F-to-remove values and thereafter selecting the appropriate predictor variables by stepwise regression or by any test based method. Our procedure selects those cutoff values which produces the last occurrence of the minimum PRESS.

The results strongly suggest that K-fold cross-validation can assist in choosing the cutoff values on the basis of minimum PRESS. Indeed, using a good computer program, the split between training data and test data can be repeated several times according to different values of l in K-fold cross-validation and performances on near minimum boosted PRESS values are evaluated. We also suggest that K-fold cross-validation is extremely useful and provide the best estimates if the correct value of K is chosen. It is less expansive than Leave-one-out cross-validation.

Experimental researcher has shown that optimal number for fold is 10, see Breiman et al (1984), Kohavi (1995). For large sample size the chance of K is in the domain $[1, 10]$ is largely arbitrary. We have proposed a procedure by introducing alternative estimates of boosted prediction sum of squares to choose the appropriate tuning parameter that is K in K-fold cross-validation for each data set. From the above results we conclude that our approach of applying K-fold

cross-validation to optimize cutoff values in stepwise regression thus very helpful in fitting the accurate predictive regression model. I hope that more interesting results will follow on further exploration of the K-fold cross-validation resampling technique.

References

- Anderssen E., Dyrstad K., Westad F., and Martens H., Reducing Over-Optimism in Variable Selection by Cross-Model Validation. *Chemometrics and Intelligent Lab. Systems* 84 (2006) 69-74.
- Aldrin M., Improved Prediction Penalizing Both Slope and Curvature in Additive Models. *Computational Statistics and Data Analysis* 50 (2006) 267-284.
- Baumann K., and Stiefl N., Validation Tools for Variable Subset Regression. *J. of Comp.-Aided Molecular Design*. 18(2004) 549-562.
- Burman P., A comparative Study of Ordinary Cross-Validation, V-fold Cross-Validation and the Repeated Learning Testing Methods. *Biometrika*. 76 (1989) 503-514.
- Burman P., Estimation of Optimal Transformation Using v-fold Cross-Validation and Repeated Learning –Testing Methods. *Sankhya Ser. A* 52 (1990) 314-345.
- Breiman L., Friedman J.H., Olshen R.A., and Stone, Classification and Regression Tree. *Wadsworth, Belmont, CA* (1984).
- Breiman, L., Better Subset Regression Using the Nonnegative Garrote. *Technometrics* 37 (1995) 373-384
- Bengio Y., and Grandvalet Y., No Unbiased Estimator of the Variance of K-fold Cross-Validation. *Journal of Machine Learning Research* 5(2004) 1089-1105.
- Clark R.D., Boosted Leave-Many-Out Cross-Validation: The Effect of Training and Test Set Diversity on PLS Statistics. *J. of Comp.-Aided Molecular Design*. 17(2003) 265-275.
- Duan K. et. Al., Evaluating of simple Performance Measures for Tuning SVM Hyperparameters. *Neurocomputing* 51(2003) 41-59.

- Mahmood and Khan: K-Fold Cross-Validation for Predictive Models in Stepwise Regression
- Davison, A.C., Hinkley, D.V., *Bootstrap Method and Their Applications*. Cambridge University Press, Cambridge, UK, 1997.
- Dudoit S. and Van der Laan M.J., Asymptotics of Cross-Validated Risk Estimation in Estimator Selection and Performance Assessment. *Statistical Methodology* 2(2) (2005) 131-154.
- Efron B., How Biased is the apparent Error Rate of a Prediction Rule? *J. Amer. Statist. Assoc.*, 81 (1986) 461-470.
- Geisser, S., The Predictive Sample Reuse Method with Applications, *J. Amer. Statist. Assoc.*, 70 (1975) 320-328.
- Glantz, S.A. and Sliker, B.K., *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, 2000.
- Herzberg A.M., and Tsukanove A.V., A Note on the Modification of Jackknife Criterion for Model Selection. *Utilitas Math.* 29 (1986) 209-216.
- Jonathan P., Krzanowski and McCarthy, On the Use of Cross-Validation to Assess Performance in Multivariate Prediction. *Statistics and Computing*. 10 (2000) 209-229.
- Kittler J., Feature Set Search Algorithm. In *C.H. Chen, Editor, Pattern Recognition and Signal Processing*, Sijthoff & Noordhoff, (1978) 41-60.
- Kohavi R., A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint conf. on Artificial Intelligence (IJCAI). 1995.
- Kutner, M. H., Nachtsheim, C. J., and Neter, J., *Applied linear Regression Models* (4th Editoin) McGraw-Hell/Irvin Series, 2004.
- Lachenbruch P.A., and Mickey M.R., Estimation of Error Rates in Discriminant Analysis. *Technometrics*. 10 (1968) 1-11
- Pudil P., Novovicova J., and Kittler J., Floating Search Method in Feature Selection. *Pattern Recognition Letters*, 15(11) (1994) 1119-1125

- R Development Core Team. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; URL <http://www.R-project.org>, (2004).
- Snee, R. D., Validation of regression models: methods and examples. *Technometrics*, 19 (1977) 415-428.
- Stone, M., Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, 36 (1974) 111-147.
- Salahuddin and Hawkes, A.G., Cross-Validation in Stepwise Regression. *Comm. Statist. Theory Meth.*, 20(4) (1991) 1163-1182.
- Shao J., Linear Model Selection by Cross-Validation. *Am. Stat. Assoc.* 88(422) (1993) 486-494.
- So S.S., Heldon S.P., Geerestein and Karplus M., Quantitative Structure- Activity Relationship Studies of Progesterone Receptor Binding Steroids. *J. Chem. Inf. Comput. Sci.* 40 (2000) 762-772
- Stearns S.D., On Selecting Features for Pattern Classifiers. In *Proc. Of the 3rd Int. Joint Conf. on Pattern recognition*, Coronado, CA, USA, (1976) 71-75.
- Somol P. and Pudil P., Oscillating Search Algorithms for Feature Selection. In *Proc. Of the 15th Int. Conf. on Pattern recognition (ICPR'2000)*, IEEE Computer Society (2000) 406-409.
- Somol P., Pudil P., Novovicova J., and Paclik P., Adaptive Floating Search Method in feature Selection. *Pattern Recognition Letters*, 20(11-13) (1999) 1157-1163.
- Vehtari A., and Lampinen J., Model Selection via Predictive Explanatory Power. *Helsinki Uni. Of Tech, Lab of Computational Engg. Publications*. Report B. ISSN (2004) 1455-1474.
- Van der Laan M.J., Dudoit S., and Kales S., Asymptotic Optimality of Likelihood-Based Cross-Validation. *Stat. Appl. Genet. Mol. Biol.* 3 (1), Art. 4, (2004) pp. (electronic)

Mahmood and Khan: K-Fold Cross-Validation for Predictive Models in Stepwise Regression

Van der Vaart , Aad W., Dudoit, S., and van der Laan, M.J., Oracle Inequalities for Multi-Fold Cross-Validation. *Statistics and Decisions*, 24, No. 3, (2006) 351–371

Wisnowski, J.W., Simpson, J.R., Montgomery, D.C., Runger, G.C., Resampling Method for Variable Selection in Robust Regression. *Comp. Stats. & Data Analysis* 43 (2003) 341-355.