Pak. J. Statist.
2016 Vol. 32(1), 21-36

IMPROVED RESAMPLING TECHNIQUE FOR THE CHOICE OF STOPPING CRITERION AND MODEL SELECTION IN STEPWISE LOGISTIC REGRESSION

Zafar Mahmoood[1], **Salahuddin**[2,3] and **Peter Salzman**[4]

[1] Department of Mathematics, Statistics and Computer Science, The University of Agriculture, Peshawar, Pakistan. Email: zafarjan@gmail.com

[2] Department of FAMCO, University of Dammam, Dammam, Kingdom of Saudi Arabia.

[3] Institute of Management and Information Sciences, CECOS University of IT and Emerging Sciences, Peshawar, Pakistan. Email: salahuddin_90@yahoo.com

[3] Department of Biostatistics and Computational Biology, The University of Rochester, USA. Email: psalzman@bst.rochester.edu

ABSTRACT

In recent years, logistic regression models are widely used in many research fields for establishing relation between a discrete outcome variable and predictor variable(s). Various automated selection procedures are used for the selection of predictor variables that might influence the outcome variable. The significance level $\left(\chi^2_{(\alpha)}\right)$ is the standard stopping criterion in these automated model selection methods in logistic regression. The problem with these automated model selection methods is the choice of appropriate stopping criterion for entry and removal of predictor variables. Most of the statistical packages typically used the default significance value $(\alpha = 0.05)$ for an entry that may be unreasonable and sometimes even dangerous because it results either too many variables in the model for a reliable interpretation or too few variables for best prediction. Besides different recommendations concerning the entry and removal criteria, there is still a problem for the true best choice of these values in automated selection methods in logistic regression models. We propose to resolve this problem by using cross-validation resampling technique that will optimize the stopping criterion each time for different data sets and different number of significant predictor variables. We further proposed the bootstrap resampling screening test for validating the final parsimonious logistic regression model. Moreover, for moderate correlated predictor variables, our strategy had shown better results as compared to other model selection methods.

KEYWORDS

Resampling; Stopping Criteria; Stepwise Logistic Regression; Variable Selection.

1. INTRODUCTION

Regression analysis has become an integral component of any data analysis and research for establishing relationship between outcome and predictor variables, fitting of statistical models and validating the predictions. In regression analysis, it is often the case when the outcome variable is discrete, taking on two (dichotomous) or more (polychotomous) possible values. Logistic regression is a model building technique and has become the standard method over the last decade to cope with such cases. We consider the general logistic regression model,

$$Log\left(P\left(Y=1/P\left(Y=0\right)\right)\right)=logit\left(P\right)=log\ \left(P/\left(1-P\right)\right)$$
$$=\beta_0+\sum_{i=1}^{k}\beta_i X_i,\quad i=1,\ 2,.....,k\ .$$

where Y is the outcome variable and X_i are the *i*th predictor variables.

Model selection is the fundamental task in regression analysis when we have a large number of predictor variables. A common problem is to select those predictor variables in the final regression model that might influence the outcome (response) variable. It is worth mentioning that if we have 10 predictors, the number of all possible models is $2^{10}=1024$. With 20 predictors, we have more than 1,000,000 possible models and with 30 predictors the number of possible models is greater than 1,000,000,000. Thus, even with moderate number of predictors, it is wise to apply automated model selection methods for selecting the subset model.

A most parsimonious, yet biologically reasonable model is one that fit the data adequately. But it should serve as an accurate predictor for new observations and should not be complex. The researchers have the problem of selecting the relevant predictors that should enter the final regression model. Various automatic model selection techniques are available in the literature. These are forward selection, backward elimination, stepwise regression and best subset selection procedures with different optimization criteria, such as Mallow's C_p, Akaike information criterion (AIC) or the Bayesian Information Criterion (BIC) etc. that are commonly used. The first three methods are based on the same idea and we will talk only about stepwise selection in logistic regression, as it is more flexible and sophisticated selection procedure.

In stepwise logistic regression method, variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. The algorithm used to define these procedures in logistic regression is almost similar as discussed for linear regression. The stepwise linear regression applies F-test, since the error is assumed to be normally distributed. While, in logistic regression the error is assumed to follow a binomial distribution and its significance is assessed via the likelihood ratio chi-square test. Thus, at any step in the procedure the most important variable, in statistical term, is the one that produces the greater change in log-likelihood relative to a model not containing the variable (that is, the one that would result in the largest likelihood ratio statistic, denoted commonly by G).

2. COMPUTATIONAL PROCEDURE FOR STEPWISE LOGISTIC REGRESSION

In logistic regression, the likelihood ratio test (G_M - statistic) is analogous to the F-statistic in linear regression (Hosmer and Lemeshow, 2000; Menard, 2002). The stepwise logistic computation is based on the calculation of log-likelihood, the likelihood ratio test (G_M - statistic) and the respective probability values of these statistic's. It is updated from step to step by applying the general algorithm and our R code for the entry and removal of predictor variables. The following formal procedure by our algorithm is applied:

1. Input the observed data for 'n' observations of 'p' predictor variables and a response variable say "class". α_{in} and α_{out} (also called P_E and P_R) are the significance levels or the probability values for entering and removing predictor variables from the logistic regression model. The probability value for removal is usually greater than or equal to the probability value for the entry.
2. Initialize the logistic regression model for NULL case, also called intercept only logistic model and compute the log-likelihood value of the model.
3. Find the predictor variable from those not in the model that has the largest G_M-statistic when added to the logistic regression model. If its p-value is at least as small as a pre-specified value, α_1 then add the variable to the model. Similarly, check for other predictor variable in turn to enter into the logistic regression model. Stop if no variable can be added.
4. Find the predictor variable between those in the model obtained in step 4 that has the smallest G_M-statistic when removed from the logistic regression model. If its p-value is greater than a pre-specified value, α_2 then delete the variable from the model. Repeat this step until no further variables can be deleted; then go to step 3.
5. The process terminates when no other predictor variable has p-value as smaller as a pre specified value (α_1) or greater than a pre specified value (α_2).

3. STANDARD STOPPING CRITERION) IN FORWARD, BACKWARD AND STEPWISE LOGISTIC REGRESSION

The most difficult and crucial aspect of using stepwise logistic regression is the choice of appropriate stopping criterion $\left(P_E \ \& \ P_R \text{ or } \alpha_{in} \ \& \ \alpha_{out}\right)$ for entry and removal of predictor variables. The level of significance α can vary from $\alpha = 0$ (no predictor variable entered in the logistic model) and $\alpha = 1$ (all predictor variables are in the logistic model). Recommendations relating to these stopping criteria were given in the literature. The usual conventional value $\alpha_{in} = \alpha_{out} = \alpha = 0.5$ has been commonly used but it has not clarified that whether it is the best criterion for predicting in the stepwise logistic model. Bendel and Affifi (1977) compared stopping criterion in forward stepwise and recommend that entry criterion between 0.15 and 0.25 is small enough to keep noise variables from being included in the model and large enough to allow authentic variables to enter to the model. Kennedy and Bancroft (1971) reported that $\alpha = 0.15$ gives the best overall results and suggested that it might be applicable to the stepwise selection algorithm. Hoerl et al. (1986) recommended that the retention criterion should be set at a

value equal to one half of the entry criterion. Draper and Smith (1988) and Flack and Chang (1987) used entry and removal criterion of 0.15 in their investigation of stepwise procedure. Lee and Koval (1997) reported in their simulation that the best α for the entry criterion in forward stepwise logistic regression is 0.05 and 0.40. Their overall recommendation is that $0.15 \le \alpha \le 0.20$ should be used for the $\chi^2_{(\alpha)}$ stopping criterion. Hosmer and Lemeshow (2000) used $P_E = 0.15$ for entry and $P_R = 0.20$ for removal, but also criticized that for broader goal of analysis $P_E = 0.25$ or even larger might be reasonable choice.

In addition, different statistical packages used different default stopping criteria. For example, SPSS uses default values for entry and retention 0.05 and 0.10 respectively (see, Norusis, 1985). SAS procedure uses $\alpha_{in} = \alpha_{out} = \alpha = 0.05$ for forward, backward and stepwise algorithms(SAS Inst., 1985). BMDP also uses values of 0.05 for both entry and retention criteria(Dixon et al., 1988). Beside these different recommendations, we still have problems for optimizing the stopping criteria in stepwise logistic regression. Hosmer and Lmeshow (2000) also stated that P_R must exceed P_E to guard against the possibility of having the program, enter and remove the same variable in successive steps. To overcome the problem in choosing the true stopping criteria, we apply cross-validation as a method to optimize P_R and P_E on the bases of the minimum stable average PRESS.

4. THE METHOD

We compare the usual conventional stopping criteria with those optimized by our method using cross-validation in the stepwise logistic model. The basic idea behind the selection of these optimal stopping criteria (P_E and P_R values) is the one we have applied in stepwise regression for optimizing cutoff values $\left(F_{in} \text{ and } F_{out}\right)$ in linear regression and also the one proposed by Lee and Koval (1997) in terms of the estimated true error rate of a prediction. The method is based upon the following steps.

Step 1: Read input data consisting of n-rows (observations) and m-columns (p-predictor variables and a response variable).

Step 2: Initialize the leave-one-out cross-validation (LOOCV) by deleting the first observation or leave-k-out cross-validation (LKOCV) by deleting a first Kth sub-sample (group) of observations. Also initiate values or sequence of α_{in} and α_{out} values.

Step 3: Run forward, backward or stepwise logistic regression and predict the deleted case(s) by using the selected model and compute the residual.

Step 4: Return to step 2 and repeat the whole procedure by deleting instead observation 2, then observation 3 and so on in LOOCV until all the observation have been deleted once or deleting instead sub-sample 2, sub-sample 3 and so on in LKOCV until all the sub-sample's have been deleted once.

Step 5: Change the α_{in} and α_{out} values or sequence of α_{in} and α_{out} values by some appropriate increment and return to step 2.

Step 6: Compute L_1 or L_2 values for PRESS. Where L_1 is the average sum of the absolute difference between actual and predicted values and L_2 is the average sum of the squared deviation between actual and predicted values.

A more natural way to express predictive accuracy is by means of expected absolute departures of observed from predicted outcome values, that is, by the absolute prediction error, measured on a scale familiar to the investigator.

Under strictly normal y, the standard deviation has an intuitive interpretation (for example, approximately two-thirds of the observed outcomes are within ±1 standard deviation of the expected outcome value) and also is a more efficient estimate than the absolute prediction error. However, both advantages were lost already under very mild departures from normality (see Hampel, 1998). Therefore, we preferred absolute prediction error for a unified concept of predictive accuracy, applicable to various outcome distributions. Our program applies L_2 prediction error for PRESS although it can be switched to L_1.

Step 7: The procedure continues for all the α_{in} and α_{out} values or sequence of α_{in} and α_{out} values defined in the vector of the cross-validation function.

Step 8: Select that value of α_{in} and α_{out} which produces the minimum L_1 or L_2 for PRESS. In practice, the minimum values of L_1 or L_2 are produced for a range of values of α_{in} and α_{out}. Here, we apply the principle of parsimony and select that value of α_{in} which produce the first occurrence (smallest of α_{in}) of the minimum L_1 or L_2 values for PRESS. Because the smallest α_{in} that is the smallest significance level of stepwise logistic regression fit a smaller number of predictor variables in the final model for the full data.

Step 9: Finally, we apply the stepwise logistic regression to the full data set on the basis of these optimized α_{in} and α_{out} stopping criteria.

Our computer program in R for implementing the above procedure containing different functions which can be efficiently applied for forward selection (FS), backward elimination (BE), stepwise (SW) selection using LOOCV and/or LKOCV, is available from the author on demand.

5. THE SCREENING STEP FOR OUR FINAL MODEL SELECTION

The bootstrap method originally proposed by Efron (1979) is used to improve our model selection procedure and validate the regression model. In the logistic regression setting, the proposed bootstrap technique is as follows.

i) Assume that our generated data or the original data is denoted by; class, $V_i,\ i=1,2,\ldots,p$, where class is the response variable and V_i are the *i*th predictor variables or covariates.
ii) We randomly select a sample of size 'n' with replacement from the original data set to obtain a bootstrap sample. Repeat the procedure large number, say B times, to obtain our bootstrap replications.
iii) Run stepwise regression model while using our optimized stopping criterion (α_{in} and α_{out}) by LOOCV or LKOCV procedure.
iv) Compute the percentage of inclusion (PI) as the number of times each predictor variable appears in the model over the total number of bootstrap replication.

The basic idea of our bootstrap diagnostic stepwise selection is that an automated model selection method for logistic regression (or any generalized regression model) is conducted to choose the prognostic important variable(s) at each bootstrap replication. In this study, we used forward stepwise selection with the optimum selection criterion obtained by our method that allows us to find parsimonious regression model. Also, we defined PI, the percent inclusion of a predictor variable over the total number of bootstrap replications and record it for the subset model selected. A predictor variable with no or smaller prognostic influence will have low PI value, since we assumed that each bootstrap replication is a random sample from the original data distribution and thus should reflect the underlying structure of the given data. Accordingly, the PI is then used as a criterion for the prognostic influence and importance of the predictor variable. We select only those variables whose PI $\geq$ 30%. A predictor variable with PI < 30% indicates no relationship with the outcome variable, a predictor variable having percent includes 30% $\leq$ PI $\leq$ 70% is called a week factor and a predictor variable with PI $\geq$ 70% is called a strong factor for inclusion. A useful discussion about PI cutoff points can be found in Sauerbrei and Schumacher (1992). The methodology, how the bootstrap stepwise procedure works is given in the table below.

Table 1
Bootstrap Screening Results for 5 Bootstrap Replications

bi/Vi	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	*				*		*			
2					*		*			*
3			*	*	*				*	*
4			*	*	*	*				*
5	*			*	*		*			*
PI (%)	40	0	40	60	100	20	60	0	20	80

'*****' indicates that the predictor variable is included in the model selected by the bootstrap resampling method.

6. SIMULATION STUDIES

In this section, several simulations were conducted to assess the adequacy of choosing the appropriate stopping criterion (α_{in} and α_{out}) for sequential selection of predictor variables in the final logistic regression model and the success rate of our implemented procedure. To implement the method described above, it is desirable to have as artificial data generated for selecting the known significant predictor variables. We made a small general program for generating any number of observations (n -rows) and any number of variables (m-columns) showing p-predictor variables with known true predictor variables and the last column is taken as the response variable named as 'class' have values 0 or 1.

Altogether, three simulation studies are presented here, each consisting of 100 runs. In each run, we generate $n=100$ observations with $p=10$ predictor variables. Half of the observations (nObs1=50) are labeled zero that is failure in response $(Y=0)$ and half successes for response $(Y=1)$. In the first simulation, we generated all non-predictive variables that is no true predictor variable(s) (nTrueVars=0). In the second and third simulation, nTrueVars = 1, 2 respectively. The effect of the variables in the true model is chosen as to be rather small that is effecting Size = 1.

We compared four variable selection algorithms: classical stepwise (CSW) with the default stopping criterion, stepwise AIC, stepwise BIC and our stepwise method after optimizing the appropriate stopping criterion, we called it optimized stepwise (OSW). 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 R which can be efficiently used for optimizing cutoff values and then selecting the final regression model.

Table 2
Simulation Study: Success Rate Based on 100 Repetitions

Method	Simulation		
	I	II	III
CSW	0.48	0.50	0.56
StepAIC	0.19	0.23	0.26
StepSBC	0.63	0.61	0.60
OSW	0.76	0.78	0.74

Table 2 report the success rate (based on 100 trials) that each method is able to find the correct subset model with the given number of true variables (0, 1 and 2) for our three simulated data sets. We see that OSW is better at selecting the right predictor variables with a significantly high probability then using CSW or AIC, BIC criterion based methods.

Moreover, to examine the effect of the selection procedure on the basis of our selected stopping criterion for the data when correlation exists between the predictor variables on the implemented procedures, several simulations were conducted. We generated three data sets of sample size $n = 50,100,500$ observations each with $p = 10$ predictor variables. Each data set is replicated 100 times with the number of true predictor variables nTrueVars = 0, 1, 3, 5 and 7 and the correlation pattern between predictor variables is 0.00, 0.05, 0.10, 0.20 and 0.50.

Table 3
Simulation Study: Success Rate Based on 100 Repetitions for Correlated Data

nVar and nTrueVar	Method	Sample size														
		50					100					500				
		Correlation					Correlation					Correlation				
		.00	.05	.10	.20	.50	.00	.05	.10	.20	.50	.00	.05	.10	.20	.50
nVar=10 nTrue=0	CSW	0.65	0.55	0.60	0.50	0.60	0.55	0.65	0.55	0.65	0.70	0.40	0.40	0.70	0.75	0.70
	StepAIC	0.30	0.30	0.30	0.30	0.12	0.10	0.20	0.30	0.10	0.30	0.20	0.10	0.20	0.35	0.35
	StepSBC	0.60	0.60	0.55	0.50	0.65	0.60	0.75	0.75	0.60	0.75	0.79	0.78	0.79	0.80	0.78
	OSW	0.82	0.69	0.68	0.85	0.86	0.76	0.83	0.81	0.76	0.73	0.82	0.80	0.87	0.84	0.82
nVar=10 nTrue=1	CSW	0.50	0.40	0.40	0.40	0.20	0.40	0.40	0.55	0.40	0.10	0.60	0.65	0.45	0.00	0.00
	StepAIC	0.25	0.15	0.30	0.10	0.10	0.10	0.15	0.20	0.20	0.00	0.30	0.20	0.30	0.00	0.00
	StepSBC	0.55	0.40	0.44	0.40	0.35	0.44	0.60	0.60	0.40	0.00	0.80	0.75	0.05	0.00	0.00
	OSW	0.97	0.80	0.85	0.82	0.55	0.80	0.93	0.80	0.79	0.30	0.93	0.75	0.65	0.35	0.00
nVar=10 nTrue=3	CSW	0.46	0.51	0.45	0.23	0.20	0.35	0.65	0.40	0.35	0.00	0.70	0.30	0.00	0.00	0.00
	StepAIC	0.16	0.15	0.25	0.10	0.00	0.30	0.45	0.15	0.06	0.00	0.35	0.16	0.00	0.00	0.00
	StepSBC	0.55	0.45	0.59	0.21	0.10	0.65	0.67	0.60	0.40	0.00	0.80	0.75	0.05	0.00	0.00
	OSW	0.85	0.90	0.85	0.80	0.45	0.98	0.95	0.62	0.55	0.97	0.95	0.62	0.23	0.12	0.00
nVar=10 nTrue=5	CSW	0.55	0.50	0.50	0.35	0.15	0.55	0.60	0.45	0.10	0.00	0.70	0.10	0.00	0.00	0.00
	StepAIC	0.40	0.15	0.25	0.10	0.00	0.30	0.30	0.20	0.05	0.00	0.60	0.10	0.00	0.00	0.00
	StepSBC	0.45	0.40	0.40	0.40	0.10	0.75	0.70	0.55	0.30	0.00	0.82	0.40	0.15	0.00	0.00
	OSW	0.70	0.85	0.75	0.80	0.50	0.95	0.90	0.85	0.65	0.20	0.97	0.67	0.20	0.15	0.00
nVar=10 nTrue=7	CSW	0.55	0.75	0.75	0.30	0.10	0.75	0.65	0.30	0.25	0.00	0.75	0.45	0.45	0.00	0.00
	StepAIC	0.45	0.40	0.40	0.25	0.05	0.60	0.45	0.25	0.10	0.00	0.50	0.20	0.05	0.00	0.00
	StepSBC	0.45	0.70	0.70	0.35	0.15	0.65	0.70	0.40	0.35	0.10	0.84	0.80	0.35	0.00	0.00
	OSW	0.95	1.00	0.95	0.65	0.70	1.00	0.90	0.85	0.60	0.25	0.87	0.85	0.65	0.30	0.00

The simulation results in Table 3 reports the results obtained from 100 repetitions and show the summary results of success rate for stepAIC, stepBIC, CSW and our proposed strategy OSW. The study confirms the problem with the model selection procedures when high correlated input variables are present in the data set. The probability of success rate for high correlated input variables is smaller and thus the percentage of noise

variable selected in the regression model is high. However, our method shows better results for moderately correlated input variables.

For bootstrap screening test and validation of our final regression model, similar to our previous simulation, we generated a binary variable Y_j and a multivariate normally distributed data $\left(V_{1j},\ V_{2j},\ldots,\ V_{pj}\right)$, $j=1,2,\ldots,\ n$ with some pre-specified number of true predictor variables in the designed experiment. In the simulation, we set $p=10$ predictor variables having 1, 3, 5 and 7 true predictor variables. Sample sized were chosen to be 100, 500 and 1000 and 500 bootstrap replications were performed for each sample size.

One Authentic (true) predictor variable: In the bootstrap stepwise procedure for our proposed strategy, the average PI of V_1 was 95.2%, 100% and 100% for the sample sizes 100, 500 and 1000 respectively, showing strong relationship with the outcome variable and the average PI for all the remaining noise variables are much smaller showing no relationship with the outcome variable. For AIC based stepwise method, besides the larger value of PI for variable V_1, there were other noise predictor variables with $PI \geq 30\%$ showing a week or strong relationship with the outcome variables. A comparison between automated model selection methods based on AIC and our proposed stepwise strategy (OSW) using an optimum choice of stopping criterion when V_1 is the only true predictor variable in the designed experiment, is given in Table 4. R indicates the relationship with the outcome variable.

Table 4
Comparison between Variable Selection Methods

Variables	Sample Size											
	100				500				1000			
	PI (%) and Relationship (R)				PI (%) and Relationship (R)				PI (%) and Relationship (R)			
	AIC	R	OSW	R	AIC	R	OSW	R	AIC	R	OSW	R
V_1	99.1	Strong	95.2	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_2	28.8	-	3.2	-	30.8	Weak	4.2	-	66.2	Weak	17.4	-
V_3	18.8	-	1.0	-	41.6	Weak	5.0	-	18.2	-	0.6	-
V_4	18.4	-	1.8	-	63.4	Weak	21.2	-	15.0	-	1.2	-
V_5	17.4	-	1.2	-	19.8	-	1.2	-	43.8	Weak	9.8	-
V_6	19.0	-	1.2	-	19.0	-	1.6	-	18.0	-	1.0	-
V_7	25.0	-	2.2	-	21.8	-	2.2	-	36.8	Weak	6.0	-
V_8	72.6	Strong	11.2	-	20.6	-	4.6	-	15.4	-	1.0	-
V_9	27.0	-	1.4	-	40.0	Weak	6.6	-	18.6	-	1.6	-
V_{10}	43.8	Weak	7.8	-	22.0	_	1.4	-	26.6	-	2.4	-

Three Authentic (true) predictor variables: In the bootstrap stepwise procedure for our proposed strategy, the average PI of V_1, V_2 and V_3 were 87.2%, 71.8% and 94.8% for sample of size 100, 100%, 100% and 100% for sample of size 500 and 100%, 100% and 100% for a sample of size 1000 showing strong relationship with the outcome variable while the average PI for all the remaining noise variables are much smaller showing no relationship with the outcome variable. For AIC based stepwise method, besides the larger values of PI for predictor variables V_1, V_2 and V_3, there are other noise predictor variables with $PI \geq 30\%$ showing relationship with the outcome variables. A comparison between automated model selection methods based on AIC and our proposed stepwise strategy using an optimum choice of stopping criterion with V_1, V_2 and V_3, the true predictor variables in the designed experiment, is given in Table 5.

Table 5
Comparison between Variable Selection Methods

Variables	Sample Size											
	100				500				1000			
	PI (%) and Relationship (R)				PI (%) and Relationship (R)				PI (%) and Relationship (R)			
	AIC	R	OSW	R	AIC	R	OSW	R	AIC	R	OSW	R
V_1	99.2	Strong	87.2	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_2	92.0	Strong	71.8	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_3	99.4	Strong	94.8	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_4	18.8	-	1.2	-	21.4	-	2.4	-	18.2	-	1.6	-
V_5	21.4	-	2.0	-	48.6	Weak	11.6	-	21.6	-	2.6	-
V_6	17.4	-	1.8	-	59.6	Weak	15.6	-	33.6	Weak	5.4	-
V_7	35.2	Weak	5.4	-	88.8	Strong	24.4	-	20.0	-	1.0	-
V_8	30.8	Weak	3.2	-	17.0	-	0.8	-	52.8	Weak	12.6	-
V_9	44.0	Weak	7.8	-	19.6	-	2.2	-	21.8	-	2.8	-
V_{10}	64.8	Weak	20.2	-	21.8	-	2.2	-	53.0	Weak	12.0	-

Five Authentic (true) predictor variables: In the bootstrap stepwise procedure for our proposed strategy, the average PI of V_1, V_2, V_3, V_4 and V_5 were 87.4%, 79.4%, 86.8%, 87.2% and 67.0% for sample of size 100, 100.0%, 99.6%, 98.8%, 100.0% and 100.0% for sample of size 500 and 100.0%, 100.0%, 100.0%, 100.0% and 100.0% for a sample of size 1000 and is selected as strong factors while the average PI for all the remaining noise variables are much smaller showing no relationship with the outcome variable. For AIC and BIC based stepwise method, besides the larger values of PI for variables V_1, V_2, V_3, V_4 and V_5 there are other noise predictor variables with $PI \geq 30\%$ showing strong and weak relationship with the outcome variables. A comparison between automated model selection methods based on AIC and our proposed stepwise strategy

using an optimum choice of stopping criterion with V_1, V_2, V_3, V_4 and V_5 the true predictor variables in the designed experiment, is given in Table 6.

Table 6
Comparison between Variable Selection Methods

Variables	Sample Size											
	100				500				1000			
	PI (%) and Relationship (R)				PI (%) and Relationship (R)				PI (%) and Relationship (R)			
	AIC	R	OSW	R	AIC	R	OSW	R	AIC	R	OSW	R
V_1	93.6	Strong	87.4	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_2	94.4	Strong	79.8	Strong	100.0	Strong	99.6	Strong	100.0	Strong	100.0	Strong
V_3	94.1	Strong	86.8	Strong	99.8	Strong	98.8	Strong	100.0	Strong	100.0	Strong
V_4	96.6	Strong	87.2	Strong	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_5	84.0	Strong	67.0	Week	100.0	Strong	100.0	Strong	100.0	Strong	100.0	Strong
V_6	38.8	Weak	7.4	-	42.4	Weak	8.2	-	52.6	Weak	12.4	-
V_7	30.0	Weak	5.4	-	19.0	-	1.4	-	55.6	Weak	14.0	-
V_8	45.8	Weak	11.0	-	33.8	Weak	4.4	-	48.8	Weak	7.8	-
V_9	44.8	Weak	12.2	-	33.4	Weak	2.2	-	12.4	-	0.8	-
V_{10}	72.2	Strong	21.8	-	52.8	Weak	16.2	-	30.4	Weak	7.8	-

7. APPLICATION TO REAL DATA SETS

Example 1:

To demonstrate the proposed method we used the data given by Kutner et al. (2004). This data set is concerned with a local health clinic sent flyers to its clients to encourage everyone, but especially older person at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who receives flu shot was coded $Y = 1$ and a client who didn't receive a flu shot was coded $Y = 0$. In addition, data were collected on their age $(X1)$ and their health awareness $(X2)$. The latter data were combined into a health awareness index, for which higher values indicate greater awareness. Also included in the data was client gender $(X3)$ male = 1 and female = 0. Table 7 lists the PRESS values corresponding to various α_{in} and α_{out} values and the plot of the average PRESS values, α_{in} and α_{out} are shown in Figure 1. The final model results are summarized in Table 8.

Table 7
The Stopping Criterion and their Corresponding Average PRESS Values

α_{in}	α_{out}	**Ave.PRESS**	α_{in}	α_{out}	**Ave.PRESS**	α_{in}	α_{out}	**Ave.PRESS**
0.01	0.01	0.1180148	0.14	0.15	0.1080421	0.27	0.31	0.1111853
0.01	0.02	0.1180148	0.14	0.19	0.1080421	0.28	0.28	0.1111717
0.01	0.06	0.1180148	0.15	0.15	0.1080421	0.28	0.29	0.1111717
0.02	0.02	0.1117560	0.15	0.16	0.1080421	0.28	0.33	0.1111717
0.02	0.03	0.1117560	0.15	0.20	0.1080421	0.29	0.29	0.1111717
0.02	0.07	0.1117560	0.16	0.16	0.1080421	0.29	0.30	0.1111717
0.03	0.03	0.1080421	0.16	0.17	0.1080421	0.29	0.34	0.1111717
0.03	0.04	0.1080421	0.16	0.21	0.1080421	0.30	0.30	0.1117704
0.03	0.08	0.1080421	0.17	0.17	0.1080421	0.30	0.31	0.1117704
0.04	0.04	0.1080421	0.17	0.18	0.1080421	0.30	0.35	0.1117704
0.04	0.05	0.1080421	0.17	0.22	0.1080421	0.31	0.31	0.1117704
0.04	0.09	0.1080421	0.18	0.18	0.1080421	0.31	0.32	0.1117704
0.05	0.05	0.1080421	0.18	0.19	0.1080421	0.31	0.36	0.1117704
0.05	0.06	0.1080421	0.18	0.23	0.1080421	0.32	0.32	0.1117704
0.05	0.10	0.1080421	0.19	0.19	0.1080421	0.32	0.33	0.1117704
0.06	0.06	0.1080421	0.19	0.20	0.1080421	0.32	0.37	0.1117704
0.06	0.07	0.1080421	0.19	0.24	0.1080421	0.33	0.33	0.1117704
0.06	0.11	0.1080421	0.20	0.20	0.1097927	0.33	0.34	0.1117704
0.07	0.07	0.1080421	0.20	0.21	0.1097927	0.33	0.38	0.1117704
0.07	0.08	0.1080421	0.20	0.25	0.1097927	0.34	0.34	0.1121920
0.07	0.12	0.1080421	0.21	0.21	0.1097927	0.34	0.34	0.1121920
0.08	0.08	0.1080421	0.21	0.22	0.1097927	0.34	0.35	0.1121920
0.08	0.09	0.1080421	0.21	0.26	0.1097927	0.35	0.40	0.1121920
0.08	0.13	0.1080421	0.22	0.22	0.1097927	0.35	0.35	0.1121920
0.09	0.09	0.1080421	0.22	0.23	0.1097927	0.35	0.36	0.1121920
0.09	0.10	0.1080421	0.22	0.27	0.1097927	0.36	0.41	0.1121920
0.09	0.14	0.1080421	0.23	0.23	0.1111853	0.36	0.36	0.1121920
0.10	0.10	0.1080421	0.23	0.24	0.1111853	0.36	0.37	0.1121920
0.10	0.11	0.1080421	0.23	0.28	0.1111853	0.37	0.37	0.1121920
0.10	0.15	0.1080421	0.24	0.24	0.1111853	0.37	0.38	0.1121920
0.11	0.11	0.1080421	0.24	0.25	0.1111853	0.37	0.42	0.1121920
0.11	0.12	0.1080421	0.24	0.29	0.1111853	0.38	0.38	0.1121920
0.11	0.16	0.1080421	0.25	0.25	0.1111853	0.38	0.39	0.1121920
0.12	0.12	0.1080421	0.25	0.26	0.1111853	0.38	0.43	0.1121920
0.12	0.13	0.1080421	0.25	0.30	0.1111853	0.39	0.39	0.1121920
0.12	0.17	0.1080421	0.26	0.26	0.1111853	0.39	0.40	0.1121920
0.13	0.13	0.1080421	0.26	0.27	0.1111853	0.39	0.44	0.1121920
0.13	0.14	0.1080421	0.26	0.31	0.1111853	0.40	0.40	0.1121920
0.13	0.18	0.1080421	0.27	0.27	0.1111853	0.40	0.41	0.1121920
0.14	0.14	0.1080421	0.27	0.28	0.1111853	0.40	0.45	0.1121920

Table 8
Summary Statistics for the Final Model of Local Health Clinic Data

$min.Ave.PRESS value
[1] 0.1080421
$min.alpha

α_{in}	α_{out}	Press	Model
0.03	0.03	0.1080421	2, 1
0.03	0.04	0.1080421	2, 1
0.03	0.08	0.1080421	2, 1

Coefficients:	(Intercept)	V2	V1
	-1.45778	-0.09547	0.07787

Degrees of Freedom: 158 Total (i.e. Null); 156 Residual
Null Deviance: 134.9
Residual Deviance: 105.8 AIC: 111.8

We are using this data to illustrate our proposed procedure. Our program selects the cutoff values α_{in} =0.03 and α_{out} = 0.03, 0.04, 0.08 corresponding to the foremost occurrence of the minimum average PRESS = 0.1080421 and thereafter using these stopping criteria to fit the final logistic regression model. We can use an equal stopping criteria that is $\alpha_{in} = \alpha_{out} = \alpha_0 = 0.03$ for both entry and removal of predictor variables corresponding to the minimum average PRESS. But some people prefer to make α_{out} slightly greater than α_{in} to introduce a small bias to keep a predictor variable in the model once it has been entered (Glantz and Slinker, 2000). We suggest selecting the final model on the basis of unequal optimized stopping criterion for entry and removal of predictor variables with a difference of 0.05 corresponding to the foremost occurrence of the minimum average PRESS. The idea behind the foremost occurrence of the minimum average PRESS for a range of cutoff values is that we are interested in selecting those stopping criteria that have a 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 foremost occurrence of the minimum PRESS because this will occur with the larger stopping criterion 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 Chatterjee and Hadi (2006). This data are regarded detecting ailing financial and business establishments are an important function of audit and control. Systematic failure to do audits and control can lead to grave consequences, such as the savings-and-loan fiasco of the 1980s in the United States. The data set gives some of the operating, financial ratios of 33 firms that went bankrupt after 2 years and 33 that remained solvent during the same period.

Table 9
The Stopping Criterion and their Corresponding Average PRESS Values

α_{in}	α_{out}	**Ave.PRESS**	α_{in}	α_{out}	**Ave.PRESS**	α_{in}	α_{out}	**Ave.PRESS**
0.01	0.01	0.05441945	0.14	0.15	0.04379781	0.27	0.31	0.03336245
0.01	0.02	0.05441945	0.14	0.19	0.04379781	0.28	0.28	0.03336245
0.01	0.06	0.05441945	0.15	0.15	0.04379781	0.28	0.29	0.03336245
0.02	0.02	0.05313842	0.15	0.16	0.04379781	0.28	0.33	0.03336245
0.02	0.03	0.05313842	0.15	0.20	0.04379781	0.29	0.29	0.03336245
0.02	0.07	0.05313842	0.16	0.16	0.04379781	0.29	0.30	0.03336245
0.03	0.03	0.05804511	0.16	0.17	0.04379781	0.29	0.34	0.03336245
0.03	0.04	0.05804511	0.16	0.21	0.04379781	0.30	0.30	0.03336245
0.03	0.08	0.05804511	0.17	0.17	0.04379781	0.30	0.31	0.03336245
0.04	0.04	0.05804511	0.17	0.18	0.04379781	0.30	0.35	0.03336245
0.04	0.05	0.05804511	0.17	0.22	0.04379781	0.31	0.31	0.03336245
0.04	0.09	0.05804511	0.18	0.18	0.04379781	0.31	0.32	0.03336245
0.05	0.05	0.04742108	0.18	0.19	0.04379781	0.31	0.36	0.03336245
0.05	0.06	0.04742108	0.18	0.23	0.04379781	0.32	0.32	0.03336245
0.05	0.10	0.04742108	0.19	0.19	0.04379781	0.32	0.33	0.03336245
0.06	0.06	0.04736699	0.19	0.20	0.04379781	0.32	0.37	0.03336245
0.06	0.07	0.04736699	0.19	0.24	0.04379781	0.33	0.33	0.03336245
0.06	0.11	0.04736699	0.20	0.20	0.04379781	0.33	0.34	0.03336245
0.07	0.07	0.04736699	0.20	0.21	0.04379781	0.33	0.38	0.03336245
0.07	0.08	0.04736699	0.20	0.25	0.04379781	0.34	0.34	0.03336245
0.07	0.12	0.04736699	0.21	0.21	0.04379781	0.34	0.34	0.03336245
0.08	0.08	0.04621178	0.21	0.22	0.04379781	0.34	0.35	0.03336245
0.08	0.09	0.04621178	0.21	0.26	0.04379781	0.35	0.40	0.03336245
0.08	0.13	0.04621178	0.22	0.22	0.04379781	0.35	0.35	0.03336245
0.09	0.09	0.04621178	0.22	0.23	0.04379781	0.35	0.36	0.03336245
0.09	0.10	0.04621178	0.22	0.27	0.04379781	0.36	0.41	0.03336245
0.09	0.14	0.04621178	0.23	0.23	0.04379781	0.36	0.36	0.03336245
0.10	0.10	0.04379781	0.23	0.24	0.04379781	0.36	0.37	0.03336245
0.10	0.11	0.04379781	0.23	0.28	0.04379781	0.37	0.37	0.03336245
0.10	0.15	0.04379781	0.24	0.24	0.04379781	0.37	0.38	0.03336245
0.11	0.11	0.04379781	0.24	0.25	0.04379781	0.37	0.42	0.03336245
0.11	0.12	0.04379781	0.24	0.29	0.04379781	0.38	0.38	0.03336245
0.11	0.16	0.04379781	0.25	0.25	0.04379781	0.38	0.39	0.03336245
0.12	0.12	0.04379781	0.25	0.26	0.04379781	0.38	0.43	0.03336245
0.12	0.13	0.04379781	0.25	0.30	0.04379781	0.39	0.39	0.03336245
0.12	0.17	0.04379781	0.26	0.26	0.04379781	0.39	0.40	0.03336245
0.13	0.13	0.04379781	0.26	0.27	0.04379781	0.39	0.44	0.03336245
0.13	0.14	0.04379781	0.26	0.31	0.04379781	0.40	0.40	0.03336245
0.13	0.18	0.04379781	0.27	0.27	0.03336245	0.40	0.41	0.03336245
0.14	0.14	0.04379781	0.27	0.28	0.03336245	0.40	0.45	0.03336245

Table 10
Summary Statistics for the Final Model of Surgical Unit Data

$min.ave.PRESS value
[1] 0.03336245
$min.alpha

	α_{in}	α_{out}	Press	Model
	0.27	0.27	0.03336245	1, 2, 3
	0.27	0.28	0.03336245	1, 2, 3
	0.27	0.32	0.03336245	1, 2, 3
Coefficients:	(Intercept)	V2	V1	V3
	-10.1535	0.3312	0.1809	5.0875

Degrees of Freedom: 65 Total (i.e. Null); 62 Residual
Null Deviance: 91.5
Residual Deviance: 5.813 AIC: 13.81

8. CONCLUSION

Our proposed strategy OSW is a worthy competitor for selecting important predictor variables using automated model selection methods in logistic regression model. In simulation studies and thereafter bootstrap screening test, OSW selects the correct model with significantly higher probability than using AIC, BIC and CSW automated model selection methods. Most importantly, our procedure selects stopping criterion (α_{in} and α_{out} values) for each data set on the basis of minimum prediction ability that help in choosing the parsimonious regression model with best possible prediction. In this study, we demonstrate the choice of best α_{in} and α_{out} values for the $\chi^2_{(\alpha)}$ stopping criterion for the purpose of variable selection and model prediction. Our study suggests that for each dataset, significance level α_{in} and α_{out} should beset that corresponds to the minimum prediction error and then using these selected α_{in} and α_{out} values to choose, the important predictor variables by any automated model selection method or test based methods in the final model. Our recommendation is to select variables considered in logistic regression analysis with any automated model selection method, carefully. Usually, it is wise to remove high correlated predictor variables, but for moderate correlated predictor variables, our stepwise strategy has shown better results as compared to other model selection methods.

Further simulation studies should be carried out to prove whether other procedures for variable selection in logistic regression model are more suitable, such as LASSO, shrinkage or the approach given by Azen et al. (2001).

REFERENCES

1. Azen, R., Budescu, D.V. and Reiser, B. (2001). Criticality of Predictors in Multiple Regression. *British J. Math. Stat. Psy.*, 54, 201-225.
2. Bendel, R.B. and Afifi, A.A. (1977). Comparison of Stopping Rule in Forward Stepwise Regression. *J. Amer. Statist. Assoc.*, 72, 46-53.

3. Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example* (4th Edition). John Wiley & Sons.
4. Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis* (2nd Edition). John Wiley, New York.
5. Dixon, W.J., Brown, M.B., Engelman, L., Hill, M.A. and Jennrich, R.I. (1988). *BMDP Statistical Software Manual*. Vol. 1. Berkeley: University of California Press.
6. Efron B. (1979). Bootstrap Methods: Another look at the Jackknife. *Annal. Stat.,* 9, 1-26.
7. Flack, V.F. and Chang, P.C. (1987). Frequency of Selecting Noise Variables in Subset Regression Analysis: A Simulation Study. *The American Statistician,* 41(1), 84-86.
8. Glantz, S.A. and Sliker, B.K. (2000). *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill.
9. Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*, (2nd Edition). A Wiley-Interscience Publication, John Wiley & Sons Inc., New York.
10. Hoerl, F.E., Schuenemeyer, J.H. and Hoerl, A.E. (1986). A Simulation of Biased Estimation and Subset Regression Techniques. *Technometrics*, 28, 369-380.
11. Hampel, F. (1998). Mean deviation: In *Encyclopedia of Biostatistics*, Volume 3, Armitage, P. and Colton, T. (Edition). Wiley: New York, 2488-2489.
12. Kennedy, W.J. and Bancroft, T.A. (1971). Model-building for Prediction in Regression Based on Repeated Significance Tests: *Ann. Math*., 42, 1273-1284.
13. Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2004). Applied Linear Regression Models (4th Edition) McGraw-Hell/Irvin Series.
14. Lee, K. and Koval, J.J. (1997). Determination of the Best Significance Level in Forward Stepwise Logistic Regression Based on Repeated Significance Test: *Commun. Statist. -Simul.*, 26, 559-575.
15. Menard, S.W. (2002). *Applied Logistic Regression Analysis* (2nd Edition). Thousand Oaks, CA: Sage Publications. QASS#106.
16. Norusis, M.J. (1985). *SPSS - Advanced Statistics Guide*. New York: McGraw-Hill,
17. R-Development Core Team (2004). *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.
18. SAS Institute (1999). *SAS User's Guide: Statistics*, 5th ed. Cary NC.
19. Sauerbrei, W. and Schuhmacher, M. (1992). A Bootstrap Resampling Procedure for Model Building: Application to Cox Regression Model. *Stat. Med.*, 11, 907-916.