Australian Journal of Basic and Applied Sciences, 4(12): 5813-5822, 2010
ISSN 1991-8178

Model Selection in Logistic Regression and Performance of its Predictive Ability

[1]S.K. Sarkar, [1,2]Habshah Midi, [1]Sohel Rana

[1]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, MALAYSIA
[2]Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, MALAYSIA

Abstract: Logistic regression studies often have several covariates and asked to cull these covariates to arrive at a parsimonious model. The goal is to maximize predictive power while minimizing the number of covariates in the model. Purposeful selection of covariates does not provide efficient model in case of large number of covariates while mechanical stepwise and best subsets selection procedures still provide a useful and effective model selection tools. Even with moderate number of covariates, stepwise method allows to decrease drastically the total number of models under consideration and to produce the final model on statistical ground. In spite of criticism, stepwise logistic regression has been widely used. Best subsets approach identifies key subsets of covariates on the basis of information criteria and provides predictive model. It is evident that with reasonable number of covariates, best subsets approach is a superior alternative to stepwise logistic regression to opt the predictive model.

Key words: binary response, best subsets, likelihood ratio test, information criteria, C-index, Somers' D

INTRODUCTION

In variety of regression applications, the outcome variable of interest has only two possible qualitative responses, and therefore can be represented by a binary response variable taking on values 0 and 1. Situations where the response is a dichotomous variable are quite common in social science studies and occur extensively in statistical applications. The non-iterative weighted least squares techniques and discriminant function analyses are often inappropriate for a bounded response because both will produce fitted values outside the permitted range and sensitive to the assumptions of normality (Hosmer *et al*., 1983). Logistic regression is a superior alternative to other statistical modeling techniques in case of dichotomous response variable. Less stringent logistic regression modeling approach became a common and popular technique among the researcher for describing how a binary response variable is associated with a set of explanatory variables.

Model selection is a fundamental task in data analysis, widely recognized as central to good inference. It is one of the most pervasive problems in statistical applications. Two aspects of modeling in practice are for the purpose of reliable interpretation of selected covariates on outcome variable and for prediction of the fitted model. Classical methods for model selection have not had much success in case of large number of predictor variables. There may remain several covariates that seem substantial association with the outcome but often it is unlikely that all of these are important. The components of the predictive model are obviously variables and their selection techniques are fundamental in statistical modeling because they seek to simultaneously reduce the chances of data over-fitting and to minimize the effects of omission bias. Variable selection is an area of study concerned with the strategies for selecting one subset out of a pool of explanatory variables that is able to explain or predict the response variable well enough, such that all contributions from the variables that remain unselected may be neglected or considered pure error (Schuster, 1998).

Variable selection is an important consideration when creating logistic regression models. Variables must be selected carefully so that the model makes accurate predictions, but without over-fitting the data. Selection that is potential from a large number of covariates manually is a laborious task and can overlook important parameters. In principle, given a set of variables, all possible models should be exhaustively enumerated and evaluated by the researcher to pick the most parsimonious set of predictors. Unfortunately, the number of

Corresponding Author: S.K. Sarkar, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, MALAYSIA
E-mail: sarojeu@yahoo.com

possible models quickly grows too large so that the all possible models approach may create problems and often not feasible. Thus, it is important that variable selection be automatic and to overcome such problem of variable selection is often addressed by sequential methods that start with a set of variables and attempt to grow and shrink the set by selecting which parameter should be added or removed from the set and traditionally called stepwise selection (Kiezun *et al*., 2009).

Stepwise selection is more flexible, sophisticated and intuitively appealing technique and allows for the examination of a collection of models which might not otherwise have been examined. In fact, when the number of predictor variables is large, stepwise selection method is probably the most widely used technique in modeling applications. But stepwise selection technique has serious drawbacks and often criticized severely for instability and bias in logistic regression coefficient estimates, their standard errors and confidence interval (Harrell, 2001). In addition, a sequence of likelihood ratio test or score test or Wald test is often used to control the exclusion or inclusion of variables, but these are carried out on the same data and so there will be problems of multiple comparisons for which many correction criteria have been developed. Also it is difficult to interpret the critical values associated with these tests, since each is conditional on the previous tests of inclusion and exclusion (Rencher and Pun, 1980; Copas, 1983). Besides, the choice of significance levels of entry and removal to judge the importance of covariates in stepwise selection method is still controversial (Shtatland *et al*., 2001). In reality, there is no one supermodel which is good for all purposes, and even in the same study it might often need two types of models, one for interpretation and another for prediction. The limitations of stepwise selection procedure have inspired us to use best subset logistic regression proposed by Hosmer *et al*. (1989) and information criteria by Akaike (1974) for efficiently screening the suits of logistic regression models with parsimonious sets of predictors.

A number of years ago, Lawless and Singhal (1978) proposed a method for efficiently screening non-normal regression models and providing the basis for best subsets logistic regression. Hosmer *et al*. (1989) have shown that best subset logistic regression may be performed in a straight-forward manner using any program capable of best subsets linear regression and can be considered as an alternative to stepwise selection procedure. The procedure identifies a group of subset models that give the best values of a specified criterion without requiring the fitting of all possible subset logistic regression. A best subsets approach would allow for the identification of these competing models (Sarkar and Midi, 2009). Lawless and Singhal (1978) proposed three criteria such as likelihood ratio test, multivariable Wald test and modified likelihood ratio test should be used for comparing models based on different subsets of variables and later on shift away from their original proposal and suggested to use the information criteria. The basic idea behind the information criteria is penalizing the likelihood for the model complexity the number of covariates used in the model. The most popular and frequently used of this family are the Akaike Information Criterion (AIC). It is the measures of goodness-of-fit of an estimated statistical model. It is grounded in the concept of entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality and can be said to describe the tradeoff between bias and variance in model construction.

The goal of the current study is to illustrate few methods of model selection with parsimonious set of predictors that result in a best predictive model in logistic regression. This work provides an overview of problems in multivariable modeling of social data with special attention is given to the task of predictive model selection rather than interpretative one to enhance predictive ability which may yield a more convincing model. The data source and statistical techniques are introduced in the Section 2, the performance of the fitted model is discussed in Section 3. Finally the conclusion and recommendations are proposed in Section 4.

MATERIALS AND METHODS

The Bangladesh Demographic and Health Survey 2004 (BDHS-2004) is part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning, maternal and child health. The BDHS is a source of population and health data for policymakers and the research community. In the survey there are three types of data available for the researcher as household's, women's and men's. We have been using the women's data file in the current analysis. A total of 11,440 eligible women were furnished their responses. But in this analysis there are only 2,216 eligible women those are able to bear and desire more children are considered. The women having more than and less than two living children are not involved in the analysis. Those women who has two living children and able to bear and desire more children are only considered in the analysis during the period of global two children campaign.

The variable age of the respondent, place of residence, religion, fertility preference, level of education, working status, sex preference and desired number of children are considered in the analysis. The variable fertility preference involving responses corresponding to the question, would you like to have (a/another) child? The responses are coded 0 for ‘no more’ and 1 for ‘have another’ is considered as desire for children which is the binary response variable (Y) in the analysis. The age of the respondent (X_1), place of residence (X_2) is coded 0 for ‘urban’ and 1 for ‘rural’, religion (X_3) is coded 0 for ‘non-Muslim’ and 1 for ‘Muslim’, level of education (X_4) is coded 0 for ‘primary or lower level’ and 1 for ‘secondary or higher level’, working status of respondent (X_5) is coded 0 for ‘housewife’ and 1 for ‘professional’, sex preference (X_6) is coded 0 for ‘having no preference’ and 1 for ‘having preference’ and desired number of children (X_7) are considered as covariates in the logistic regression model.

In this study, two popular predictive model selection techniques namely best subsets selection and stepwise logistic regression are discussed and compared to select the most parsimonious model. Predictive performance of the selected model is enumerated by using several summary measures of fit statistics with special attention given to the likelihood ratio test, Hosmer-Lemeshow test, Akaike Information Criterion (AIC), Concordant index etc.

2.1 Formulation of Best Subsets Logistic Regression:

Consider a random pair (Y, X) where Y is a binary response variable coded as either 0 or 1, and $X' = \left(1, x_1, x_2 \cdots x_{k-1}\right)$ is a vector of k covariates having 1 accounts for an intercept term. Assuming the n independent observations indexed by i of such pair can be modeled as

$$Y_i = \theta_i\left(X\right) + \varepsilon_i \tag{1}$$

Since ε_i is Bernoulli error, the conditional expectation that the outcome is present be denoted by $E\left(Y_i = 1 \middle| X_i'\right) = \theta_i\left(X\right)$. It is evident that S-shape curve configuration has been found to be appropriate in many applications and one of which is logistic function. The conditional mean that the characteristic of the response variable is present be well approximated by the logistic function

$$\theta_i\left(X\right) = \frac{\exp\left(Z_i\right)}{1 + \exp\left(Z_i\right)} \tag{2}$$

with $Z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{(k-1)i} = X\beta$ Here X is an n × k design matrix of explanatory variables and β is a k ×1 vector of parameters. The quantity θ_i is known as the probability for the ith covariate satisfying the important requirement $0 \le \theta_i \le 1$. Then the log-odds of having Y=1 for given X is modeled as a linear function of the explanatory variables as

$$E\left(Y \middle| X\right) = \ln\left(\frac{\theta_i}{1 - \theta_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{(k-1)i} \tag{3}$$

Since observations come from Bernoulli random variables and assumed to be independent, the log-likelihood function L (β) is defined as

$$L\left(\beta\right) = \sum_{i=1}^{n} Y_i \ln\left(\frac{\theta_i}{1 - \theta_i}\right) + \sum_{i=1}^{n} \ln\left(1 - \theta_i\right) \tag{4}$$

The most commonly used method of estimating the parameters of a logistic regression model is the method of Maximum Likelihood (ML) instead of Ordinary Least Square (OLS) method, mainly for the reason that ML method based on Newton-Raphson iteratively reweighted least square algorithm becomes more popular with the researchers (Ryan, 1997). The maximum likelihood estimator $\hat{\beta}$, of the parameter vector β is obtained by differentiating L(β) with respect to β. For convenience in multiple logistic regression models, let Y denote the vector of response values, θ denote the E(Y), the likelihood equations can be written in matrix notation as

$$\frac{\partial L(\beta)}{\partial \beta} = X'(Y-\theta) \tag{5}$$

Now, theoretically putting $\frac{\partial L(\beta)}{\partial \beta} = 0$, produces $\hat{Y} = \hat{\theta}$, satisfying $X'\left(Y-\hat{Y}\right) = 0$ Iterative estimates of β as shown by Pregibon (1981) are then obtained as

$$\hat{\beta} = \left(X'WX\right)^{-1} X'WZ \tag{6}$$

where W is an n×n diagonal matrix with general element $w_i = \hat{\theta}_i\left(1-\hat{\theta}_i\right) \hat{\theta}_i$ is the estimated logistic probability and $Z = X\hat{\beta} + W^{-1}\hat{\varepsilon}$ is a vector of observations of a pseudo-dependent variable having $\hat{\varepsilon} = \left(Y-\hat{\theta}\right)$ is the vector of residuals. The general element of the pseudo-dependent variable Z_i with a case weight $w_i = \hat{\theta}_i\left(1-\hat{\theta}_i\right)$ is

$$z_i = \left(1, x_i'\right)\hat{\beta} + \frac{\left(y_i-\hat{\theta}_i\right)}{\hat{\theta}_i\left(1-\hat{\theta}_i\right)} = \hat{\beta}_0 + \sum_{j=1}^{k-1}\hat{\beta}_j x_{ij} + \frac{\left(y_i-\hat{\theta}_i\right)}{\hat{\theta}_i\left(1-\hat{\theta}_i\right)} \tag{7}$$

The representation of $\hat{\beta}$ given in (6) provides the basis for use of linear regression program. Hosmer *et al*. (1989) justified that use the values of Z_i as dependent variable, the values of X_i as covariates and w_i as case weights the estimated coefficients produced by the analysis with a linear regression program will be identically equal to the maximum likelihood estimates that obtained by logistic regression program. The basis for efficient screening of models is to use quantities available from the fit of the full k-variable model to approximate the fit of a model containing a subset of variables. The residual sum of squares [SSE (k)] obtained from the linear regression program is

$$SSE(k) = \sum_{i-1}^{n} w_i\left(z_i-\hat{z}_i\right)^2 = \sum_{i-1}^{n} \frac{\left(y_i-\hat{\theta}_i\right)^2}{\hat{\theta}_i\left(1-\hat{\theta}_i\right)} \tag{8}$$

The residual sum of squares in (8) is identical to the Pearson chi-square (χ^2) statistic from a maximum likelihood logistic regression program. It follows that the mean residual sum squares is $\hat{\sigma}^2 = \chi^2/n-k$. The estimates of the standard error of the estimated coefficient produced by the linear regression program are $\hat{\sigma}$ times the square root of the diagonal elements of the matrix $\left(X'WX\right)^{-1}$. Thus, to obtain the correct values of the standard error of maximum likelihood estimates need to divide the estimates of the standard error produced by the linear regression program by $\hat{\sigma}$ Hence, we may use any best subsets linear regression program to execute the computations for best subsets logistic regression. Again residual sum squares for the fitted model containing the subsets of p variables such that p<k is nothing but the increase in the residual sum of squares due to excluding (k-p) variables and given by the quadratic form as

$$\lambda^* = SSE(p) - SSE(k) \tag{9}$$

Here λ^* is the unconditional multivariable Wald test statistic for the hypothesis that the coefficients for the (k-p) variables not in the model are simultaneously equal to zero. In situations where there are many possible models containing different variables for the same p and different values of p. The subsets of variables selected for best models depend on the criterion chosen for best. Several criteria have been proposed to select variables of which Mallow's (1973) measure of predictive squared error C_P is regarded as one of the efficient measure to compare such competing models. In linear regression program the Mallow's criterion for a particular subset of p variables is computed as

$$C_p = \frac{SSE(p)}{SSE(k)/(n-k-1)} + 2(p+1) - n = \frac{\chi^2 + \lambda^*}{\chi^2/(n-k-1)} + 2(p+1) - n \tag{10}$$

In order to use C_P to compare different competing models there should have a referent standard. Under the null hypothesis that all the coefficients corresponding to the (k-p) variables are simultaneously equal to zero, λ^* will be distributed for sufficiently large n, as chi-square with (k-p) degrees of freedom. Under the assumption that the model is fitted correctly, the approximate expected values of χ^2 and λ^* are (n-k-1) and (k-p), respectively. Substitution of these approximate expected values into the expression (10) yields C_P =p+1.. Hence, models with C_P near (p+1) are candidates for the best model. On the other hand, if the subset of variables under consideration has excluded important covariates, λ^* will follow a non-central chi-square distribution and C_P provides value larger than (p+1). Thus, the best subsets linear regression program selects the subset that has the smallest value of C_P. Lawless and Singhal (1978) proposed criteria for comparing models based on different subsets of variable through both multivariable Wald test given in (9) and likelihood ratio test as

$$G = \left\{(-2\ln L_p) - (-2\ln L_k)\right\} \tag{11}$$

Here L_P and L_k are the likelihoods for subset of p covariates and full model, respectively. Like Wald statistic λ^*, the likelihood ratio test statistic G is also approximately distributed as chi-square with (k-p) degrees of freedom. Hauck and Donner (1977) and Jennings (1986) examined the inferential adequacy of the multivariable Wald statistic in logistic regression and justified that λ^* behaved in an aberrant manner often failing to reject the null hypothesis when the predictor was significant. The implication is that λ^* and hence C_P may be small for subsets of variables whose coefficients are not all equal to zero. To overcome such problem, Lawless and Singhal (1987) modified their earlier idea and proposed another measure of the badness of a statistical model with parameters determined by the method of maximum likelihood defined by Akaike (1974) and known as Akaike Information Criterion (AIC) for p covariates as

$$AIC_p = -2Log_e[L(\hat{\beta})] + 2(p+1) \tag{12}$$

AIC not only rewards goodness-of-fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over-fitting and preferred model is the one with the lowest AIC value. Table 1 exhibits the five best models selected using C_P as the main criterion. The values of λ^*, likelihood ratio test G for the variable excluded from the model, the corresponding degrees of freedom with p-values and AIC_P are also included in Table 1.

Table 1: Five best models identified using Mallow's C_p and AIC_p. Model covariates, Mallow's C_p, AIC_p, Wald statistic λ^*, likelihood ratio test G for the excluded covariates with degrees of freedom and p-value

Model	Model Covariates	C_P	AIC_P	λ*	G	dt	P-value
1	X_1, X_6, X_7	5.79	1989.04	6.44	6.52	4	0.16
2	X_1, X_5, X_6, X_7	5.40	1988.33	3.78	3.81	3	0.28
3	X_1, X_2 X_5, X_6, X_7	4.68	1987.28	0.75	0.75	2	0.69
4	X_1, X_4 X_5, X_6, X_7	6.37	1989.18	2.64	2.65	2	0.27
5	X_1, X_3 X_5, X_6, X_7	7.31	1990.23	3.68	3.71	2	0.16

In Table 1, the test statistic G and λ* have similar values, as expected, since they test the same hypothesis under the same asymptotic distribution. Using summary statistics of different criteria, we should select model 3 as the best model since it has the smallest value of C_P and AIC_P. The Wald and likelihood ratio tests with high p-values also suggest that the excluded variables are not significant. It is important to note that best subsets selection procedure identified four of the five models as best which included age of the respondent (X_1), working status (X_5), sex preference (X_6) and desired number of children (X_7) as common. Thus, these variables are obviously important and should be in any model. Among the four, third model having the minimum values of C_P and AIC_P which contains important social variable place of residence (X_2) and provide the optimal predictive model under study.

2.2 Stepwise Model Selection Approach:

In practice, there are times when the outcome being studied is relatively new and the important covariates may not be known and associations with the outcome not well understood. In such instances, most studies collect many possible covariates and screen them for significant associations. Suppose the pool of potential covariates contains 10 or more variables, use of a best subsets algorithm may not be feasible. An automatic search procedure that develops the best subset of covariates sequentially may then be helpful. Employing a stepwise selection procedure can provide a fast and effective means to screen a large number of predictors and to fit a number of logistic regression equations simultaneously. The main goal of modeling is to maximize predictive power of the model while minimizing the number of predictors included in the model. Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the importance of variables, and either includes or excludes them on the basis of a fixed decision rule. The importance of a variable is defined in terms of a measure of the statistical significance of the coefficient for the variable. The statistic used depends on the assumptions of the model. In logistic regression the errors are assumed to follow a Bernoulli distribution, and significance is assessed via the likelihood ratio chi-square test. Thus, at any step in the procedure the most important variable, in statistical terms, is the one that produces the greatest change in the log-likelihood relative to a model not containing the variable.

The main controversy of stepwise logistic regression is the choice of entry and removal level for inclusion and exclusion of the predictors to judge their order of importance. Literature suggests the choice of .05 as entry and removal level is too stringent, often excluding important variables from the model. Hosmer and Lemeshow (2000) strongly recommended using 0.15 as entry level and 0.20 as removal level for stepwise logistic regression program to include the important predictors which are associated with the response variable. Several statistical tests have been proposed to evaluate the significance of the entry or removal covariates in the model under stepwise selection procedure as the likelihood ratio test, the score test and the Wald test. Given a choice, Moulton et al. (1993) suggested that the likelihood ratio test is a superior alternative to test the significance of each predictor because research has shown it has the optimal statistical properties. Under this technique, the criterion for entry is based on a test of the significance of the coefficient for the predictor X_i conditional on (X_1, X_2,... X_{i-1}) being in the model. Here we may test the conditional null hypothesis H_0: $\beta_i = 0$ such that (X_1, X_2,... X_{i-1}) remains prior in the model, against H_1: $\beta_i = 0$. To test the null hypothesis the test statistic is

$$G = \left\{\left(-2\ln L_{i-1}\right) - \left(-2\ln L_i\right)\right\} \tag{13}$$

Here L_1 and L_{i-1} are the maximized log-likelihoods for the step i and (i-1), respectively. Under the null hypothesis, G is approximately distributed as chi-square with 1 degree of freedom. Table 2 reflects the selected variables by stepwise method using maximum likelihood ratio test defined in Equation (13). In Table 2 we may compare the p-value for entry at each step to the pre chosen entry level of significance. The p-values corresponding to the likelihood ratio test until step (5) are less than 0.15 and indicate that the covariates X_7, X_1, X_6, X_2 and X_5 provide a significant contribution along with the order of importance to predict the response variable.

The p-values corresponding to the covariates X_4 and X_3 exceed 0.15 which indicate that they do not provide any significant advantage to predict the model efficiently. Thus, the computer output for the forward selection procedure terminates at step (5), because no further predictors can be added with resulting p-values less than 0.15. Hence, the final model would be the one with all covariates entered until step (5). The final model identified using stepwise selection is the same as that identified earlier by best subsets selection procedure.

Table 2: Sequential analysis of likelihood ratio test by forward stepwise method using 0.15 as entry and 0.20 as removal levels

Step (*i*)	Covariate entered	$2\text{In}L_i$	*G*	*dt*	*p-value*
0	Intercept	2761.26	-	-	-
1	X_7	2094.37	666.89	1	0.000
2	X_1	1992.93	101.44	1	0.000
3	X_6	1981.04	11.89	1	0.001
4	X_2	1978.03	3.01	1	0.082
5	X_5	1975.28	2.75	1	0.097
6	X_4	1974.64	0.64	1	0.424
7	X_3	1974.52	0.12	1	0.729

3. Predictive Power of the Selected Model:

Once selection of the potential covariates has been completed, a multivariable model be constructed and the importance of each variable included in the model should be verified by an examination of the Wald statistic for each variable. A logistic regression model is fitted to the data using the selected covariates and results are presented in Table 3. The output from the current study suggested the change in coefficient estimates from the sequential analysis were not substantial. The critical evaluation of the individual predictor in Table 3 reveals that the selected covariates in the final model are significantly associated to the response variable.

The predictive power of the selected model can be determined by assessing few summary measures of fit. In fact, summary measures of goodness-of-fit, as they are routinely provided as computer output with the fitted model, give an overall indication of the fit of the model. The commonly employed summary measures of goodness-of-fit are presented in Table 4. Under the global null hypothesis, the selected covariates have no significant contribution to predict the outcome variable Pearson chi-square, deviance chi-square, maximum likelihood ratio chi-square and score chi-square tests measure how well they affect the response variable. These tests indicate that the overall logistic regression model with the selected covariates is superior to the null model and suggest that the model is highly significant and may adequately be used for prediction. In the Hosmer-Lemeshow (2000) goodness-of-fit test, the subjects are divided into approximately ten groups of roughly the same size based on the percentile of the estimated logistic probabilities. The discrepancies between the observed and the expected frequencies in these groups are summarized by the Pearson chi-square statistic, which is then compared to the chi-square distribution with 8 degrees of freedom. In this test, poor fit is indicated by significance value less than 0.05. To support the predictive model, a significance value larger than 0.05 is needed. The high p-value 0.729 signifies that there is no substantial difference between the observed and predicted responses and indicates that the model predicts the data very well.

Table 3: Critical evaluation of Maximum Likelihood Estimates using selected covariates

Covariates	$\hat{\beta}$	$S.E\left(\hat{\beta}\right)$	$W=\hat{\beta}^2/\left[S.E\left(\hat{\beta}\right)\right]^2$	*dt*	*p*-value	$\exp\left(\hat{\beta}\right)$
Constant	-3.15	0.40	62.02	1	0.000	0.04
X_1	-0.09	0.01	81.00	1	0.000	0.91
X_2	0.21	0.12	3.06	1	0.080	1.23
X_5	-0.23	0.14	2.70	1	0.100	0.79
X_6	0.59	0.17	12.04	1	0.001	1.80
X_7	1.98	0.14	200.02	1	0.000	7.24
Log - Likelihood = - 987.64		G = 785.98	df = 5	*p*-value < 0.00		

There exists no simple satisfactory measure to test the significance of the fitted logistic regression model like R^2 in linear regression. Cox and Snell (1989) also proposed one such measure in logistic regression as $R^2=1-\left\{L(0)/L\left(\hat{\beta}\right)\right\}^{2/n}$, where $L(0)$ and $L\left(\hat{\beta}\right)$ is the likelihood for the null model and the saturated model, respectively. An alternative form proposed by Nagelkerke (1991) is defined as

$\bar{R}^2 = R^2/\max\left(R^2\right)$. The maximum possible value of R^2 is $\max R^2 = 1-\left\{L\left(0\right)\right\}^{2/n} = 0.75$. The values of these statistics are presented in Table 4 which indicates that the model performs at an acceptable level.

Table 4: Summary measures of goodness-of-fit statistics of the model with selected covariates

Summary Statistic	Value	*dt*	p-value
Pearson χ^2	2457.51	2208	0.000
Deviance χ^2	1974.52	2208	0.000
Likelihood ratio χ^2	785.98	5	0.000
Score χ^2	732.93	5	0.000
Hosmer- Lemeshow χ^2	5.28	8	0.729
Cox and Snell R^2	0.30		
Nagelkerke R^2	0.42		
Predicted *CCR*	82.80		
Area under ROC curve	0.82		
Somers' rank correlation D_{xy}	0.64		

Another supplementary measure to assess the worth of the fitted model is known as Correct Classification Rate (CCR). From the fitted logit model, we may calculate the fitted probabilities for each observation. If the fitted probability for an observation is greater than or equal to 0.5 and less than 0.5, we may classify it to 1 and 0, respectively. Then we determine what proportion of the data is classified correctly. A high proportion of correct classification will signify that the logistic model is working well. A low proportion of correct classification will indicate poor performance. It is suggested that the proportion of observation classified correctly by the logistic regression should be much higher than the base (cut value 0.5) level for the logistic model to be deemed. The predicted CCR for the current model is 82.8. Based on this measure, we may also conclude that the model performance is good.

A unit less index of the strength of the rank correlation between predicted probability of response and actual response is a more interpretable measure of the fitted model's predictive discrimination. One such index is the probability of concordance C between predicted probability and response usually known as C-index. The C-index, which is derived from the Wilcoxon-Mann-Whitney two sample rank test, is computed by taking all possible pairs of subjects such that one subject responded and other did not. The index is the proportion of such pairs with the responder having a higher predicted probability of response than the non responder. The predictive performance of the model can be measured by calculating C-index. Traditionally C-index is identical to the area under Receiver Operating Characteristic (ROC) curve (Hanley and McNeil, 1982). The area under the ROC curve is a useful summary measure of the model's predictive power. The predictive ability of the model can be quantified by area under ROC curve. This curve plots the probability of correctly classifying a positive response (Sensitivity) against the probability of incorrectly classifying a negative response (1-specificity) for the entire set of possible cut-off points. The area under the ROC curve, which ranges from 0 to 1, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not. When area under ROC curve is 0.5, it is thought of as predicting at random with no discrimination and values close to 1 indicate that the model has outstanding predictive ability. As a general rule, if area under ROC curve lies within, $0.8 \le ROC < 0.9$, the predictive ability of the model is excellent (Hosmer and Lemeshow, 2000). Using the rule, the area under ROC curve for the current model is 0.82 and hence its predictive ability is highly satisfactory.

A similar index of measuring predictive accuracy known as Somer's rank correlation coefficient between predicted probabilities and observed outcomes defined as $D_{xy} = 2(C-0.5)$, where C is the concordant index or area under ROC curve. In case of $D_{xy} = 0$, the model is making random prediction with no discrimination and $D_{xy} = 1$, indicates the model discriminates perfectly. $D_{xy} = 0.64$, presented in Table 4 for the current model indicates the acceptable capability of prediction performance.

4. Discussion and Conclusion:

The common approach to statistical modeling involves seeking the most parsimonious model that still predicts the data. The criteria for inclusion of a variable in the model vary between problems and disciplines. The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more easily generalized. The problem of popular purposeful univariable selection approach is that it ignores the possibility that a collection of variables, each of which is weakly associated with the outcome, can become an important predictor of outcome when taken together (Hosmer and Lemeshow, 2000).

To overcome the situation, the subset of variable selection approach has been developed. Best subsets selection is a technique that has not been used extensively in logistic regression. With this approach, a number of models containing one, two, three variables and, so on are examined to determine which are considered the best according to some specified criteria one of which is AIC. AIC is asymptotically equivalent to the cross-validation and bootstrapping criteria which is based on predictive ideas. Moreover, AIC is considered a cornerstone of the modern approach to prediction (Stone, 1977). Thus best subsets selection procedure select predictive model. The main advantage of best subsets selection is that many more models can be quickly screened than was possible with the other approaches to variable identification. One potential disadvantage with this approach is that one must be able to fit the model containing all the possible covariates and for a large number of covariates, selection of the parsimonious set may not feasible (Hosmer *et al*., 1989).

Another approach to variable selection is to use a stepwise method which identifies variables automatically as candidates for a model solely on statistical grounds. One problem in stepwise logistic regression is that arbitrary criteria are used to arrive at the stepwise model. This arbitrariness adds much of uncertainty to the selection process and can reduce substantially the degree of automation in stepwise selection. With arbitrary and unreliable selection criteria, entirely different variable sets will be selected by different modelers and by different samples. Unfortunately, there has been no better alternative that overcome these problems and still gives a parsimonious model. Thus most social scientists still using stepwise logistic regression.

Stepwise, best subsets and other mechanical selection procedures have been criticized because they can select noise covariates. It can be compromised only when the analyst understands the strengths, and especially the limitations, of the methods that these methods can serve as useful tools in the model building process. The analyst, not the computer, is ultimately responsible for the review and evaluation of the model. So, analysts should not lured into accepting the covariates suggested by these mechanical strategies without considerable critical evaluation.

In the current study, the covariates identified using stepwise selection is the same as those identified earlier by best subsets selection procedure. But this may not always be the case. In such a situation with reasonable number of covariates, it is strongly recommended to use best subsets selection approach as a superior alternative to stepwise selection method to opt the predictive model.

REFERENCES

Akaike, H., 1974. A new look at a statistical model identification. IEEE Transactions in Automatic Control, 19: 716-723.

Copas, J.B., 1983. Regression Prediction and Shrinkage. Journal of Royal Statistical Society, B45: 311-354.

Cox, D.R. and E.J. Snell, 1989. The Analysis of Binary Data, 2nd edition. Chapman and Hall, London.

Hanley, J.A. and B.J. McNeil, 1982. The measure and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143: 29-36.

Harrel, F.E., 2001. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York, Inc.

Hauck, W.W. and A. Donner, 1977. Wald's Test as applied to hypotheses in logit analysis. Journal of the American Statistical Association, 72: 851-853.

Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression, 2nd edition. Wiley, New York.

Hosmer, D.W., B. Jovanovic and S. Lemeshow, 1989. Best Subsets Logistic Regression. Biometrics, 45: 1265-1270.

Hosmer, T., D.W. Hosmer and L.L. Fisher, 1983. A comparison of the maximum likelihood and discriminant function estimators of the coefficients of the logistic regression model for mixed continuous and discrete variables. Communications in Statistics, B12: 577-593.

Jennings, D.E., 1986. Judging inference adequacy in logistic regression. Journal of the American Statistical Association, 81: 471-476.

Kiezun, A., T.A. Lee and N. Shomron, 2009. Evaluation of optimization techniques for variable selection in logistic regression applied to diagnosis of myocardial infraction. Bioinformation, 3(7): 311-313.

Lawless, J.F. and K. Singhal, 1987. ISMOD: An all-subsets regression program for generalized linear models, I. Statistical and Computational background. Computer methods and Programs in Biomedicine, 24: 117-124.

Lawless, J.F. and K. Singhal, 1978. Efficient screening of nonnormal regression models. Biometrics, 34: 318-327.

Mallows, C.L., 1973. Some comments on C_p. Technometrics, 15: 661-676.

Moulton, L.H., L.A. Weissfeld and R.T. St.Laurent, 1993. Bartlett correction factors in logistic regression models. Computational Statistics and Data Analysis, 15: 01-11.

Nagelkerke, N.J.D., 1991. A note on the general definition of the coefficient of determination. Biometrika, 78: 691-692.

Pregibon, D., 1981. Logistic regression diagnostics. Annals of Statistics, 9: 705-724.

Rencher, A.C. and F.C. Pun, 1980. Inflation of R^2 in Best Subset Regression. Technometrics, 22: 49-54.

Ryan, T.P., 1997. Modern Regression Methods. John Wiley & Sons, Inc.

Sarkar, S.K. and H. Midi, 2009. Optimization Techniques for Variable Selection in Binary Logistic Regression Model Applied to Desire for Children Data. Journal of Mathematics and Statistics, 5(4): 387-394.

Schuster, C., 1998. Regression Analysis for Social Sciences. Academic Press, New York.

Shtatland, E.S., E. Cain and M.B. Barton, 2001. The Perils of Stepwise Logistic Regression and how to escape them using Information Criteria and the Output Delivery System, SUGI'26, Proceeding, Paper, 222-26, Cary, NC: SAS Institute, Inc.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society, B39: 44-47.