International Journal of Applied and Natural Sciences (IJANS)
ISSN (P): 2319-4014; ISSN (E): 2319-4022
Vol. 7, Issue 2, Feb – Mar 2018; 41-52

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

VARIABLE SELECTION PROCEDURES FOR LOGISTIC REGRESSION MODELS

Adekanmbi, D.B

Research Scholar, Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

ABSTRACT

Variable selection procedures are applicable to predictive model building process such as logistic regression, and generally for generalized linear modelling. The essence of variable selection is to select the best parsimonious adequate model among the available models for a data set, to avoid using redundant predictors in a model. In this study, variable selection procedures suitable for logistic regression model are considered namely: stepwise procedures, criterion-based procedures and cross-validation procedures. The three procedures of variable selection were exemplified on predictive logistic models using real life data sets on births and coronary heart disease (CHD) to determine the most suitable variable selection procedure for the logistic regression models. The logistic regression model for the birth data is to estimate the functional relationship between the binary response variable, type-of-birth and the predictors. For the coronary heart disease (CHD) data the interest is to explore the relationship between the risk factors, such as age, sex and cholesterol level of patients and the presence or absence of CHD in the study population. The stepwise procedures were computationally intensive. The criterion-based procedures and cross-validation procedures are investigated in this study, though, involve a wider search but in a preferable manner compared to the stepwise procedures that use restricted search through the space of potential models. It is therefore recommended to use criterion-based procedures when building a predictive logistic regression model for a data set with dichotomous response variable.

KEYWORDS: *Forward Selection Method, Backward Elimination Method, Stepwise Selection, Leave- One-Outcross-Validation (LOOCV), k-Fold Method, Delete-d Cross-Validation*

Article History

Received:* *17 Dec 2017* | *Revised:* *30 Jan 2018* | *Accepted:* *05 Feb 2018

INTRODUCTION

A good generalized linear model (GLM) should obey the principle of parsimony. The principle of parsimony is to avoid over-fitting to achieve a good model fit that can predict well, [8, 11, 14]. The purpose of variable selection is to select a model as small as possible with the best subsets of predictors, which gives a good fit and predicts the dependent variable well, usually referred to as parsimonious model. A parsimonious model is the simplest model among plausible models for a phenomenon with the best subset of predictors to explain a data set, [14, 15]. Variable selection procedures are suitable for building the most parsimonious model for generalised linear models (GLM).

The aim of variable selection is to construct a model that predicts well or explains the relationship in a data set, [22, 23]. This is necessary so as to avoid redundant predictors in a model which could add noise to the estimation of other variables of interest, and thereby cause degrees of freedom to be wasted, [19, 23]. Variable selection also helps in avoiding co linearity among the predictors, and the cost of implementing a parsimonious model for prediction is reduced since unnecessary predictors must have been removed. Prior to variable selection, it is necessary to exclude outliers and influential observations from a data set, and transform any variable that seems appropriate. Model selection procedures are more stable than selecting model with the best overall average performance, [12, 23]. A natural technique to select predictors in the context of GLM is to use the common mechanical variable selection methods which are: Forward selection, Backward elimination, and Stepwise selection methods, [2, 7, 18, 22].

When the numbers of predictors considered in a GLM is large, subset selection methods may be appropriate to determine the most influential predictors. *Bestglm* package has been developed in *R* statistical computing software which uses search algorithm to find the GLM model with smallest deviance, especially when the number of predictors considered in a GLM is quite large, [18, 25]. *Bestglm* is based on using information criteria to select the best model out of large possible models. The information criteria include Akaike Information Criterion (AIC), Schwarz Information Criterion (BIC), and BIC_q. Another approach to model selection in the *bestglm* package is cross-validation. The approach includes leave-one-out (LOOCV), k-fold and delete-d cross validation, [18].

The purpose of this study is to investigate the mechanical variable selection procedure, a variety of information criteria based procedure as well as the cross-validation procedure, in order to determine the most suitable approach for generalised linear model, and specifically for logistic regression model. The different variable selection approaches will be exemplified in this study using real life data sets to determine the most suitable of these approaches to logistic regression modelling.

Stepwise Procedure

The stepwise procedure involves three approaches which are forward selected, backward elimination and stepwise selection to find the best generalized linear model.

Backward Elimination

Backward elimination takes place by removing predictor already in a generalized regression model if the predictor is not significant. The method involves starting with all predictors in a model. The predictor with the highest p-value greater than the critical value is then removed, [5, 7, 8, 23]. The model can be refitted and any other predictor with highest p-value should be removed. Once a predictor is removed by this method, it remains removed. The process continues until all the p-values for the remaining predictors in the model are less than the critical value.

Forward Selection

Forward selection method is the reverse of the backward elimination. The method involves starting with no predictor in the model. For all predictors not in the model, their p-value should be checked and the predictors with lowest p-value less than the critical value should be added to the model, [2, 12, 22]. The process continues until no new predictor can be added.

Stepwise Selection

Stepwise selection is a combination of backward elimination and forward selection methods. A forward selection step can be followed by a backward elimination step. Inclusion and deletion of predictors is done one at a time. At each stage, a predictor may be added or removed. Stepwise selection method concludes if no further predictor can be added to a model. The method however has some disadvantages. It is possible to miss the optimal model because of the one-at-time adding and dropping of predictors which may lead to instability of selection, [12, 22]. Also, the removal of redundant predictors could amplify the statistical significance of the remaining predictors. The procedure tends to select models that are smaller than desirable for prediction, and such predictions could be of worse quality than from a full model, [10, 22].

Information Criteria-based Procedure

Information criteria-based procedures are used to choose the best model out of the k+1 model cases. Given *p* potential predictors, then there are 2^p possible models, [14, 17, 18]. Information Criterion-based procedure is to find out of all 2^p subsets, the best subset based on some criteria. Some of the criteria are:

Akaike Information Criterion (AIC)

AIC is most commonly used as a selection criterion for GLM. AIC selection criterion provides the best approximating model among a candidate set of models, [1, 17].

$$AIC = -2In(L) + 2p \quad \textbf{(1)}$$

where

L: maximized value of the likelihood function for the estimated model.

p: number of parameters in the model.

Bayesian Information Criterion (BIC)

BIC as defined by [21]:

$$BIC = -2In(L) + pIn(n) \quad \textbf{(2)}$$

The model with smallest AIC or BIC is preferred. BIC penalizes larger GLM models more heavily than AIC and therefore tend to prefer smaller models compared to AIC, [2, 23]. A suitable function for this procedure in R computing software does not evaluate the AIC for all possible models, but uses a search method that compares models sequentially. The procedure has some comparison to the stepwise method, but with the advantage that no dubious p-value is used.

BIC_γ Criterion

One of the drawbacks of BIC is that the criterion tends to select models with many predictors. Chen and Chen, (2008)[2] therefore suggested a prior uniform of models instead of a prior uniform of all possible models. The general form of BIC_γ is

$$BIC_\gamma = -2\ln(L) + p\ln(n) + 2\gamma\ln\binom{k}{p} \quad \textbf{(3)}$$

where

γ : an adjustable parameter.

p: number of parameters in the model.

k: number of possible input variables without the bias or intercept term.

When $\gamma = 0$, BIC_γ reduces to BIC. When $p = 0$, BIC_γ corresponds to only intercept term, while p=k corresponds to using all parameters that are equally likely a priori.

BIC_q Criterion

Xu and Mcleod, (2010) [25] derived BIC_q criterion by assuming that each parameter has a prior Bernoulli of *q* of being included, where $q \in [0,1]$. BIC_q is therefore given as

$$BIC_q = -2\ln(L) + p\ln(n) - 2p\ln\left(\frac{q}{1-q}\right) \quad \textbf{(4)}$$

BIC_q is equivalent to BIC when $q = 1/2$. Also q=0 and q=1 are equivalent to selecting the models with p=k and p=0 respectively. An interval estimate for q that is based on confidence probability α, with $0 < \alpha < 1$ was derived by [25].

Cross-Validation Procedure

Another approach to model selection that is noteworthy is cross-validation (CV) approach. Leave-one-out cross-validation (LOOCV), K-fold and delete-d CV (D-CV) are some of the cross-validation methods. The cross-validation approach involves narrowing the field to the best models of size *p* for *p=0,1,2,…k* and then comparing each of the *k+1* possible models using cross-validation to select the best one. The model of size *p* with the smallest deviance is then chosen as the best model.

Leave-One-Out Cross-Validation (LOOCV)

In LOOCV procedure, one observation is removed, say i, and the regression is refit. The prediction error denoted as $\hat{e}_{(i)}$ for the omitted observation is computed. The process is repeated for all observations and the prediction error sum of squares is calculated as

$$PRESS = \sum_{i=1}^{n} \hat{e}_{(i)}^2 \quad \textbf{(5)}$$

The disadvantage of this method of this variable selection is that the method is not usually accurate compared to the other CV methods. The method usually has high variance, as commented by [11, 20].

K-Fold Method

With K-fold method, the data is divided randomly into K folds of roughly equal size, which forms a partition of the observations, 1,2,….,n so that the set of observations in k^{th} partition is denoted as $\prod_k$. A fold is selected as the validation sample, while the remaining partitions are used as training sample. The performance is calibrated on the validation sample, and this is repeated for each fold. The average performance over the K folds is determined. In order to make subset selection, the validation sum-of-squares is calculated for each of the K validation samples using the formula

$$S_k = \sum_{i \in \prod_k} \lim \left(\hat{e}^{(-k)_i}\right)^2 \quad \textbf{(6)}$$

where

$\hat{e}^{(-k)_i}$: the prediction error when the k^{th} validation sample is removed, and the model fit to the remaining data and then used to predict the observations $i \in \prod_k$ in the validation sample.

Also, the cross-validation score is computed as

$$cv = \frac{1}{n} \sum_{k=1}^{K} S_k \quad \textbf{(7)}$$

Where

n: number of observations.

For each validation sample, the estimate of the cross-validation mean square error may be obtained as follows:

$$cv_k = \frac{S_k}{N_k} \quad \textbf{(8)}$$

Where

N_K: number of data points in the k^{th} validation sample.

Given that s^2 is the sample variance of $cv_1, cv_2, \ldots cv_K$. An estimate of the sample variance of cv, that is the mean of $cv_1, cv_2, \ldots cv_K$ is s^2 / K.

The interval estimate for CV is therefore: $cv \pm s / \sqrt{K}$. This is an indication that the most parsimonious adequate model is the model with the best CV score in the interval. This rule improves the stability of the k-fold method greatly, [18, 20].

Delete-d CV

The method was proposed by [20]. Random samples of size *d* are used as the validation set, and many validation sets are generated in this manner, while the complementary part of the data is used each time as the training set. When *d=1*, the delete-d CV is approximately equivalent to LOOCV, and even yield the same results provided enough validation sets are used. The method becomes consistent when *d* increases with n. [20] suggested that letting $\lambda_n = \log n$, then

$$d^* = n\left(1-(\log n - 1)^{-1}\right) \quad \textbf{(9)}$$

Where

n: the number of observation.

Logistic Regression Model

The logistic regression model is suitable for modelling discrete response variable having binary or dichotomous categories, [8, 9, 11, 14, 15, 17]. The model is part of a category of statistical models called generalized linear models, and is simply referred to as model for binary responses, [15, 17, 24]. With two categories of birth categorized as single or multiple, logistic model can be used to predict which of the two categories of birth a pregnant woman is likely to have, given certain other information.

Given that Y_i's are independent binomial random variables with parameters n_i, p_i. The probability distribution function of Y_i is therefore given by [8, 14, 15, 17]:

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \text{ for } y_i = 0,1,2,\ldots..n_i \quad \textbf{(10)}$$

With $Y_i \sim B(n_i, p_i)$ under the assumption that p_i is constant, it follows that $\mu_i = E(Y_i) = n_i p_i$ so that $p_i = \frac{\mu_i}{n_i}$, and $Var(Y_i) = n_i p_i (1-p_i)$.

That logistic regression with k predictors can be written as

$$\text{logit}(p_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots\ldots + \beta_k x_{ik} \quad \textbf{(11)}$$

In matrix form,

$$\eta_i = x_i'\beta \quad \textbf{(12)}$$

where

$$x_i'\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots\ldots + \beta_k x_{ik}.$$

x : is a vector of covariates, so that $x_{i1}, x_{i2}, \ldots\ldots\ldots, x_{ik}$ are the predictors.

β : is a vector of regression coefficients. The β's are the regression coefficients associated with the *k* variables.

i: indicates individual observations.

Case I: Birth data

The purpose of the logistic regression model is to estimate the functional relationship between binary response variable, *type-of-birth* (whether single or multiple births) and the predictors, which are *age* of mothers, *parity* of mothers, *religion* of mothers, and *tribe* of mothers. The continuous predictors are *age* and *parity* of mothers, while the other predictors are categorical in nature. The logistic model for the birth data has linear predictors such that:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3^{(1)} X_{i3}^{(1)} + \beta_4^{(1)} X_{i4}^{(1)} \quad \textbf{(13)}$$

Where

X_{i1} : is the effect of *age* of mothers.

X_{i2} : is the effect of *parity* of mothers.

$X_{i3}^{(1)}$: is the effect of *religion* of mothers, fitted as a categorical variable, with one dummy variable for the 2 levels of religion.

$X_{i4}^{(1)}$: is the effect of *tribe* of mothers, also fitted as a categorical variable with one dummy variable for the 2 levels of tribe.

The regression coefficients are estimated by the maximum likelihood method which is designed to maximize the likelihood of producing the data given the parameter estimates. The link function is $g(p_i) = \eta_i = \log\left(\frac{p_i}{1-p_i}\right)$.

Case II: Coronary Heart Disease (CHD) Data

The data on coronary heart disease is extracted from Framingham study, [4, 13]; to investigate the relationship of a number of potential risk factors to the occurrence of coronary heart disease (CHD) for a sample of subjects selected to participate in the study. The risk factors to consider in the study are gender, age and cholesterol level of the patients. The response variable is the presence or absence of CHD in a patient, which is binary in nature. This study focuses on modelling the extent to which CHD is associated with predictors. The functional form of the relationship between CHD, sex, age, and cholesterol level of CHD patients is:

$$\eta_i = \beta_0 + \beta_1^{(1)} X_{i1}^{(1)} + \beta_2^{(1)} X_{i2}^{(1)} + \beta_3^{(1)} X_{i3}^{(1)} \quad \textbf{(14)}$$

Where

$X_{i1}^{(1)}$: is the effect of *sex* of patients, fitted as a categorical variable, with one dummy variable for the 2 levels of sex.

$X_{i2}^{(1)}$: is the effect of *age* of patients, fitted as a categorical variable, with one dummy variable for the 2 levels of age.

$X_{i3}^{(1)}$: is the effect of *cholesterol level* of patients, fitted as a categorical variable, with one dummy variable for the 2 levels of cholesterol.

Results of the Analysis

Case I: Results of the Analyses of the Births Data

The response variable type indicates whether a woman gives birth to either single birth or multiple births. The variable selection procedures discussed in this study will be illustrated on the data set aimed to investigate how the response variable is associated with the predictors: *age*, *religion*, *tribe* and *parity* of a woman, and build a predictive model.

The stepwise logistic regression results on the response variable type are provided in Table 1. Results of the full model, forward selection logistic regression, backward elimination logistic regression, and stepwise logistic regression are provided in table 1. Only the final fitted models are shown in the table. In this illustration, the distribution is binomial with logarithmic link function usually referred to as a logistic regression model. It is appropriate to use the binomial distribution since the response variable is dichotomous in nature.

Table 1: Summary of the Full Model, Forward, Backward and Stepwise Selection Procedures for the Birth Data

	Full Model				
Variable	**Estimate**	**Std. Error**	**z-Value**	$\Pr(>\|z\|)$	**AIC**
Intercept	-1.72965	0.70312	-2.460	0.0139*	923.5
Age	-0.01653	0.02410	-0.686	0.4926	
Religion2	-1.68766	0.89037	-1.895	0.0580	
Parity	-0.10736	0.08697	-1.234	0.2170	
Tribe2	0.96722	0.25496	3.794	0.0001 ***	
Age:Religion2	0.06589	0.03054	2.157	0.0310 *	
Forward Selection Method					
Variable	**Estimate**	**Std. Error**	**z-Value**	$\Pr(>\|z\|)$	**AIC**
Intercept	-1.72965	0.70312	-2.460	0.0139*	923.5
Age	-0.01653	0.02410	-0.686	0.4926	
Religion2	-1.68766	0.89037	-1.895	0.0580	
Parity	-0.10736	0.08697	-1.234	0.2170	
Tribe2	0.96722	0.25496	3.794	0.0001 ***	
Age:Religion2	0.06589	0.03054	2.157	0.0310 *	
Backward Elimination Method					
Variable	**Estimate**	**Std. Error**	**z-Value**	$\Pr(>\|z\|)$	**AIC**
Intercept	-1.48159	0.66967	-2.212	0.0269 *	923.05
Age	-0.02964	0.02159	-1.373	0.1699	
Religion2	-1.67666	0.88671	-1.891	0.0586	
Tribe2	0.95479	0.25441	3.753	0.0002 ***	
Age:Religion2	0.06485	0.03040	2.133	0.0329 *	
Stepwise Method					
Variable	**Estimate**	**Std. Error**	**z-Value**	$\Pr(>\|z\|)$	**AIC**
Intercept	-1.48159	0.66967	-2.212	0.0269 *	923.05
Age	-0.02964	0.02159	-1.373	0.1699	
Religion2	-1.67666	0.88671	-1.891	0.0586	
Tribe2	0.95479	0.25441	3.753	0.0002 ***	
Age:Religion2	0.06485	0.03040	2.133	0.0329 *	

Signif. codes: 0 '***' 0.001'*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

According to the Forward method, the model that includes all the four predictors and the interaction term is the best model, so that none of the variables is removed. Backward elimination performed on the data set give rise to results that are similar to the Forward procedure, except that *parity* is removed. Also the variables *tribe2* and the interaction term *age:religion2* are significant at 0.001and 0.01 levels respectively for the Forward, backward elimination and the stepwise methods. The final models for backward elimination and stepwise are virtually the same with the following predictors: *age, religion2, tribe2,* and the interactive term between *age:religion2* retained in the final model. Conclusively, the results of the Forward method are similar to the Backward elimination and the Stepwise method, but the only exception is that Forward method retains *parity* predictor, though the predictor is not significant.

Table 2: Summary of the Subset Models Based on AIC, BIC, kfold and LOOCV Criteria

Bestglm AIC					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
BICq Equivalent for q in (0.00384594715204034, 0.945072116893121)					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
Bestglm BIC Equivalent for q in (0.00384594715204034, 0.945072116893121**)**					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
BICg Equivalent for q in (0.00384594715204034, 0.945072116893121)					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
BICq Equivalent for q in (0.00384594715204034, 0.945072116893121)					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
Kfold *BICq Equivalent for q in (0.00384594715204034, 0.945072116893121)*					
Variable	**Estimate**	**Std. Error**	**t-Value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	
Bestglm(LOOCV) BICq Equivalent for q in (0.00384594715204034, 0.945072116893121)					
Variable	**Estimate**	**Std. Error**	**t-value**	$\Pr(>\lvert t\rvert)$	
Intercept	0.09090909	0.02617091	3.473669	5.361551e-04	
Tribe2	0.12671441	0.02976733	4.256828	2.275445e-05	

The results of the criterion-based logistic regression are shown in Table 2. From the results, the best model based on AIC has only one predictor which is *tribe2* and is significant. The subset models based on BIC_g, BIC_q, kfold and LOOCV are quite the same with the subset model based on AIC and BIC criteria, with *tribe2* as the only predictor in the subset models.

Case II: Results of the Analyses of the Coronary Heart Disease (CHD) Data

The response variable in the CHD data is whether a patient has developed coronary heart disease or not. The stepwise logistic regression results on the response variable type are provided in Table 3, showing the extent to which CHD is associated with the risk factors: *sex, gender,* and the *cholesterol level*. Only the result of the full model is shown in Table 3.

Table 3: Summary of the Full Model for the CHD Data

		Full Model			
Variable	**Estimate**	**Std. Error**	**z-Value**	$\Pr(>\|z\|)$	**AIC**
Intercept	-2.7161	0.0945	-28.745	2e-16(***)	2449.7
Age1	1.1624	0.1109	10.485	2e-16(***)	
Sex1	-1.0918	0.1159	-9.421	2e-16(***)	
Chol1	0.7740	0.1127	6.869	6.48e-12	

Signif. codes: '***' 0.001

The algorithms for the Forward selection method, backward elimination method and Stepwise method give rise to results that are equivalent to the full model for the CHD data. The results of the other procedures are not presented in Table 3, to avoid unnecessary repetition. This is an indication that all the risk factors considered are highly significant in explaining the development of coronary heart disease. According to the Forward, Backward and stepwise methods, the best model is the one that includes the variables *Age, Sex and Chol*. Generally, the result of the analysis shows that the higher the cholesterol level and age of a patient, the greater the chance of developing coronary heart disease. Also, the males are more likely than females to have CHD.

Table 4: Summary of the Subset Models Based on AIC Criteria

***Bestglm* AIC**				
BICq Equivalent for q in (3.2229502289205e-08, 1)				
Variable	**Estimate**	**Std. Error**	**t-Value**	**Pr(>|t|)**
Intercept	0.07643222	0.006380805	11.978461	1.323191e-32
Sex1	-0.07168440	0.007604573	-9.426487	6.395862e-21
Age1	0.09094654	0.008395736	10.832468	4.913822e-27
Chol2	0.05721433	0.008710529	6.568410	5.618838e-11

The result of the subset model based on AIC criterion is shown in Table 4. From the results, the best model based on AIC has all the three risk factors: *Sex, Age and Chol* which are all highly significant. The results of the subset models based on BIC_g, BIC_q, kfold and LOOCV are quite similar to the result of the subset model based on AIC criterion, and are therefore not shown in the table to avoid repetition.

DISCUSSIONS

This study compared some variable selection procedures used in model selection of logistic regression model. In generalised linear modelling, the usual technique to select predictors is to use stepwise procedure. Generally, variable selection methods are sensitive to influential observations and outliers. Stepwise procedure comprising of Forward, backward elimination and stepwise methods are computationally cheap compared with the criterion-based procedure and cross validation procedure, but have lots of drawbacks. Stepwise procedure has problems in the presence of collinearity, and is expensive in terms of time and cost, using a lot of paper. It is also possible to omit the optimal model due to the one-at-a-time way of adding or deleting predictors in stepwise procedure. Also, the procedure tends to choose models that are smaller than desirable for prediction purposes and also amplify the statistical significance of the predictors in the model. Some authors have criticized stepwise procedure on the ground that it could be computationally intensive.

Another selection procedure considered as available in *R* computing software includes a variety of information criteria procedure and the cross-validation procedure. The *bestglm* package in *R* computing software used for the analysis is based on exhaustive search algorithm to find the logistic model with smallest deviances. The approaches in *bestglm* package are not without disadvantages. The approaches in *bestglm* could require more computer time when the explanatory variables are more than 10. The computer timing may not be important in some data analysis, but could be a major concern when simulation is involved. Furthermore, cross-validation procedures cannot be implemented in *R* computing software when there are categorical variables present in a data set with three or more levels, unless an exhaustive enumeration approach is used. This observation is consistent with the view of Mcleod and Xu, (2010) [25] on the use of *bestglm* package in analysing generalised linear model.

CONCLUSIONS

In generalized linear modelling, it is important to conduct variable selection procedure to select the most parsimonious model, in order to avoid using redundant predictors in a model. Existing variable selection procedures were compared for logistic models on real life data sets to determine the most suitable of these procedures for logistic regression model. It was discovered that the criterion-based procedure and cross-validation procedure usually involve a wider search in a preferable manner compared to the stepwise procedure that use restricted search through the space of potential models. It is therefore preferable to use criterion-based procedures as variable selection method when analysing a data set with dichotomous response variable using logistic regression modelling.

REFERENCES

1. *Akaike, H., 1986. A new look at the statistical model identification. IEEE Trans. Autom.Contr. AC-19: 716-723.*
2. *Chen, J. and Z. Chen, 2008. Extended Bayesian information criteria for model selection with large model space. Biometrika, 95:759-771.*
3. *Collet, D., 2002. Modelling binary data. (2nd Edn.). Chapman & Hall.*
4. *D'Agostino, R.B., R.S.Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro,W.B. Kannel, 2008., "General cardiovascular risk profile for use in primary care: the Framingham Heart Study". Circulation, 117 (6): 743–753.*
5. *Dobson, A.J., 2010. An intodction to generalised linear models. CRC Press.*
6. *Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani, 2004. Least angle regression (with discussion). Annals of Statistics, 32:407-499.*
7. *Efron, B. and T. Hastie, 2016. Computer age statistical inference: algorithms, evidence and data science. New-York: Cambridge University Press.*
8. *Faraway, J.J., 2006. Extending the linear model with R: generalised linear, mixed effects and nonparametric regression models. New-York: Chapman & Hall/ CRC.*

9. *Green, P.J. and B.W. Silverman, 1993. Non-parametric regression and generalized linear models: a roughness penalty approach. London: Chapman & Hall.*
10. *Harrell Jr., F. E., 2001. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. New York, Berlin, Heidelberg: Springer Verlag.*
11. *Hastie, T., R.Tibshirani, J. Friedman, 2009. The elements of statistical learning: data mining, inference and prediction, 2nd Edn. Springer-Verlag, New-York.*
12. *Hosmer, D.W. and S. Lemeshow, 2000. Applied logistic regression. 2^{nd} Edn. New-York: Wiley and Sons, Inc.*
13. *Kannel, W.B., and T. Gordon, 1970. The Framingham Study. An Epidemiological Investigation of Cardiovascular Disease. National Institutes of Health, (Section 26).*
14. *Adekanmbi D.B, Generalized Gamma Regression Models with Application to CD4 Cell Counts Data of Aids Patients, International Journal of Applied Mathematics & Statistical Sciences (IJAMSS), Volume 6, Issue 4, June-July 2017, pp. 19-36*
15. *Krzanowski, W.J., 1998. An introduction to statistical modelling. London: Arnold.*
16. *Lawal, B., 2003. Categorical data analysis with SAS and SPSS applications. New-Jersey: Lawrence Erlbaum Associates, Inc.*
17. *Marin, J.M., and C.P. Robert, 2014. Bayesian essentials with R. 2nd Edn. New-York: Springer.*
18. *McCullagh, P., and J.A Nelder, 1989. Generalised linear models, 2nd Edn. London: Chapman andHall.*
19. *McLeod, A.I, and Y. Zhang, 2008. Improved subset autoregression: with R package. Journal of Statistical Software, 28(2):1-28. Available from: http://wwwjstatsoft.org/v28/i02. [Accessed 2 August 2017]*
20. *Molenberghs, G., and G. Verbeke, 2006. Models for discrete longitudinal data.*
21. *Shao, J., 1993. Linear model selection by cross validation. Journal of the American Statistical Association, 486-494.*
22. *Schwarz, G., 1978. Estimation the dimension of a model. Annals of Statistics, 6: 461-464.*
23. *Steyerberg, E. W., 2009. Clinical Prediction Models. A practical approach to development,validation, and updating: New-York: Springer.*
24. *Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58: 267-288.*
25. *Woodward, M., 2005. Epidemiology: study design and data analysis 2^{nd} Ed. Florida: Chapman and Hall/CRC.*
26. *Xu, C. and A.I. McLeod, 2010. Bayesian information criterion with Bernoulli prior. Submitted for publication.*