IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 13, Issue 4 Ver. I (Jul. – Aug. 2017), PP 25-33
www.iosrjournals.org

Impact of Automatic Backward Model Selection Criterion on the Bias of the Parameters in Logistic Regression Model

[1]Solomon Mulei Mutava, [2]Leonard Kiti Alii, [3].Aggrey Onago Adem
[1]*Department of Mathematics and Computer Science Pwani University,*
[2]*Department of Mathematics and Computer Science, Pwani University*
[3]*Mathematics and Physics department, Technical University of Mombasa*
Corresponding Author: Solomon Mulei Mutava

Abstract*: Logistic regression model is one of the popular mathematical models for analysis of binary data with applications in health, behavioural and statistical sciences. The main mathematical concept under the logistic regression model is the logit or the natural logarithm of an odds ratio. It is one of the most commonly used models to account for confounders in medical literature. This research sought to evaluate the impact of automatic backward model selection criterion on the bias of the parameters for large sample size. This study used a data set simulated in R-package using different kinds of controlled variables and also diabetic data obtained from Coast General Provincial hospital, Mombasa. The automatic backward selection method was used because they are faster and objective since they use the* p *-value to find the optimal model in which the fitted values are closest to the true outcome probabilities. The overall result was an inclusive logistic regression model with a subset of statistically significant predictors that best explain the variability in their observations.*

Date of Submission: 12-06-2017 Date of acceptance: 24-07-2017

I. Introduction

Logistic regression analysis technique can be used to find the best fitting model that best describes the relationship between binary outcome and the set of independent variables (Hosmer *et al*, 2013). The main mathematical concept under the logistic regression is the logit or the natural logarithm of an odds ratio. Traditionally ordinary least squares (OLS) regression or linear discriminant function analysis were used to address the analysis and prediction of a dichotomous outcome. Both techniques were subsequently found to be less than ideal for handling dichotomous outcomes due to their strict statistical assumptions namely; linearity, normality, and continuity for OLS regression and multivariate normality with equal variances and covariance's for discriminant analysis (Cabrera *et al*, 2013). Logistic regression has thus been increasingly used in social sciences, educational research especially in higher education (Peng C.Y, 2013). This prediction model is very important in clinical decision making because it can guide care providers as well as individuals in deciding further disease management.

This study seeks to evaluate the impact of automatic backward model selection criterion on the bias of the parameters for large sample size data by selecting the best variables for the logistic regression model using simulated data and a case study diabetes data from Coast General provincial hospital.

II. Literature Review

Regression methods are commonly used for analysing the relationship between dependent variable and one or more independent variables (Al-Ghamdi, 2001). The most popular regression method is linear regression using the method of least squares. It is applicable when the dependent variable is continuous, independent and identically distributed. They were the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications Yan, Xin (2009).

Traditionally ordinary least squares (OLS) regression or linear discriminant function analysis were used to address the analysis and prediction of a dichotomous outcome. Both techniques were subsequently found to be less than ideal for handling dichotomous outcomes due to their strict statistical assumptions namely; linearity, normality, and continuity for OLS regression and multivariate normality with equal variances and covariance's for discriminant analysis (Cabrera *et al*, 2013).

2.1 Logistic Regression Model

According to (Hosmer *et al*, 2000), the fact concerning the interpretability of the coefficients is the fundamental reason logistic regression has been such powerful tool for epidemiologic research. Based on its assumptions, the logit model can be estimated using the maximum log-likelihood method (Gourieroux, 2000).

According to Klein Baum *et al* (2008*)*, logistic regression quantifies the relationship between the dichotomous dependent variable and the predictors using odds ratios. Logistic regression curve is an s-shaped or sigmoid curve, often used to model population growth (Eberhardt *et al*, 2012). Logistic regression analysis technique can be used to find the best fitting model that best describes the relationship between an outcome and the set of independent variables (Hosmer *et al*, 2013). Logistic regression has thus been increasingly used in social sciences, educational research especially in higher education (Peng C.Y, 2013). The main mathematical concept under the logistic regression is the logit or the natural logarithm of an odds ratio. Logit model analyses the relationship between multiple independent variables and a categorical dependent variable and estimates the probability of occurrence of an event by fitting data to a logistic curve (Park *et al*, 2013).

2.2 Model Assumptions

McCullagh et al (1983) cited that logit model assummes the following;

i. The outcome (Y) follows an independent Bernoulli distribution
ii. A linear predictor $n_i = \beta' h(x_i) = \ln(P/(1-P))$ is the log odds ratio and P is defined as expectation of y and has a logit link function: $n_i = g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$.

Bewick *et al* (2005) cited that the following assumptions still applies to logistic regresion model:

i. The depedent variable to be discrete mostly dichotomous
ii. The desired outcome should be coded 1, model should be fitted correctly (not overfitted or underfitted with variables)
iii. The model should have little or no multicollinearity
iv. The logistic regression does not require a linear relationship between the depedent and independent variables, it requires that the independent variables are linearly related to the log odds of an event.
v. Lastly logistic regression requires large sample sizes because maximum likelihood estimates are less powerful than ordinary least squares used in linear regression.

2.3 Backward Selection

The best possible logistic regression model can be obtained by a method called stepwise backward elimination which in a predictive model is a straightforward way of reaching the highest possible accuracy (Menard, 1995). A new sample method can be used to assess the goodness of fit of a previously developed model by applying the model as it is to the new sample (Harrel *et al*, 1996). In demonstrating the generalizability of a model in order to use it to predict outcomes for future subjects, model validation has to be carried out (Hosmer *et al*, 2000). According to Hosmer *et al* (2000), full logistic regression model will have all the parameters of interest and the simple model has one variable dropped. The likelihood ratio test is chi-square distributed and if test is significant then the dropped variable will be a significant predictor in the equation (Premph, 2009)

Information criterion statistics (AIC) or Bayesian Information Criterion (BIC) ranks the evidence in the data to select good models from a set of a-priori chosen models (Burnham *et al*, 2002). Information criteria are generally preferred over multiple hypothesis tests because model building is not inherently a hypothesis testing problem and selection via hypothesis testing has shown to include unimportant variables (Burnham *et al*, 2002).

According to Hosmer and Lemeshow (2000), the deviance statistic plays an essential role in the assessment of goodness of fit of the model. A comparison between a saturated model and the current model where a saturated model is one that contains as many parameters as the number of data points and the current model (that contains only the variables being assessed) is made. Large deviance values and P-values less than 0.05 are an indication of lack of fit of the current model (Agresti, 2007). A p-value greater than 0.05 significance level is an indication that at least one coefficient is non-zero (Abdelrahman, 2010).

III. Methodology

3.1 Logistic regression

The logistic regression model is popular because the logistic function on which the model is based provides estimates that must lie in the range between 0 -1 and has the appealing S-shaped description of the combined effect of several risk factors on the risk for a disease. (David G. Kleinbaum, 2010).

The dependent variable in the logistic regression was binary or dichotomous. The maximum likelihood method, which yields values for the unknown parameters, was used for estimating the least squares function. Logistic regression solved such problems by applying the logit transformation. Logistic regression predicts the logit of Y to X. Since the logit is the natural logarithm (ln) of odds of Y and the odds are the ratios of probabilities (π) of Y happening to probabilities (1-π) of Y not happening.

The logistic regression model according to Harrell (2001) is given by equation

$$\pi(x_i) = P(y_i = 1/x_i) = \left[1 + \exp\left(-X^T \beta\right)\right]^{-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Where $y_i = \begin{cases} 1, \text{ true/pass} & \text{for } i = 1,2,\ldots\ldots,n \\ 0, \text{ Otherwise} \end{cases}$

$$X^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_{p-1} x_{p-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.2)}$$

Therefore,

$$\beta_p \times 1 = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_{p-1} \end{bmatrix}, \quad X_p \times 1 = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ X_{p-1} \end{bmatrix}, \quad X_{i_{p\times 1}} = \begin{bmatrix} 1 \\ X_{i1} \\ . \\ . \\ X_{i,p-1} \end{bmatrix}$$

where $x_{1,} x_2, \ldots\ldots\ldots x_k$ are the independent variables, β_0 is the coefficient of the constant term, $\beta_1, \beta_2, \ldots\ldots\ldots, \beta_{p-1}$ are the coefficients of the p independent variables and $\pi(x_i)$ is the probability of an event that dependents on p-independent variables.

Since $\pi(x_i) = \left[1 + \exp\left(-X^T\beta\right)\right]^{-1}$

$$= \frac{1}{1 + \exp\left(-X^T\beta\right)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.3)}$$

$$\therefore 1 - \pi(x_i) = 1 - \frac{1}{1 + \exp(-X\beta)}$$

$$= \frac{\left[1 + \exp(-X^T\beta)\right] - 1}{1 + \exp(-X^T\beta)}$$

$$= \frac{\exp(-X^T\beta)}{1 + \exp(-X^T\beta)}$$

$$\Rightarrow \frac{\pi(x_i)}{1 - \pi(x_i)} = \left[\exp(-X^T\beta)\right]^{-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.4)}$$

Hence, $\ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \log it\left[\pi(x_i)\right] = X^T\beta \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.5)}$

3.2 Model Estimation

The coefficients in logistic regression model tell us the relation between a dummy dependent variable and continuous or/and categorical independent variables. The coefficients are expected to have optimal values. This is done with the maximum likelihood estimation method which helps to find the set of parameters for which the probability of observed data is largest (Scott A., 2008).

$$\ln\left[\frac{\pi(i)}{1 - \pi(i)}\right] = -\left[-X^T\beta\right] = X^T\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots\ldots \beta_k x_{ki} \ldots \text{(3.6)}$$

Where $\pi_i = P(Y_i = 1) = 1 - P(Y_i = 0)$. From equation 3, each y_i represents a binomial count in the i[th] population, thus the maximum likelihood equation comes from the probability distribution of dependent variable Y.

The joint distribution of Y,

$$f(y/\beta) = \pi_{i=1}^{n} \frac{n_i!}{y^i!(n_i - y_i)!} p_i^{yi}(1 - p_i)^{n_i - y_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.7)}$$

The combination function $C(n_{i,}y_i)$ is the number of different ways to arrange y_i successes from n_i trials which become $\frac{n_i!}{y^i!(n_i-y_i)!}$. In any trial, the probability of success is p_i, similarly the probability of $n_i - y_i$ failures is $(1-p_i)^{n_i-y_i}$. The likelihood function is same as probability density function except that the parameters of the function are reversed. Thus the likelihood function use fixed value for Y resulting to the function;

$$f(\beta / y) = \pi_{i=1}^{n} \frac{n_i!}{y^i!(n_i-y_i)!} p_i^{yi}(1-p_i)^{n_i-y_i} \quad \text{..... (3.8)}$$

Rearranging equation 3, $\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\sum_{j=0}^{k}\beta_j x_{ji}}$. Solving for π_i and using equation 5 the equation to be maximized can be written as

$\pi_{i=1}^{n}(e^{\sum_{j=0}^{k}\beta_j x_{ji}})^{y_i} \quad (1-\frac{e^{\sum_{j=0}^{k}\beta_j x_{ji}}}{1+e^{\sum_{j=0}^{k}\beta_j x_{ji}}})n_i$. Taking logs on both sides of the equation

$$l(\beta_0,\beta_1.....\beta_k) = \ln(\beta_0,\beta_1...\beta_k) = \sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i)) \text{ .... (3.9)}$$

By simplifying and taking the derivative with respect to each β and set it equal to zero to get the critical points of the log likelihood function $\frac{\partial}{\partial\beta_j}\sum^{k}{}_{j=0}\beta_j x_{ij} = x_{ij} = \text{.....} = \text{.....} = \sum_{i=1}^{N} y_i x_{ij} - n_i p_i x_{ij} = 0$ ..... (3.10)

The estimate of β can be found by setting each of the $k+1$ equations at equations (3.10) equal to zero and solve for each β_j. This solution gives a critical point either a maximum or minimum and if the matrix of second partial derivative is negative definite it will be maximum (Scott A., 2008).

This matrix also forms the variance-covariance matrix of the parameter estimates. It can be found by differentiating each of the $k+1$ equations in equation (3.10) for the second time with respect to each β, so the second partial derivate will be of the form

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_j} = \frac{\partial l(\beta)}{\partial\beta_j}\sum_{i=1}^{N} y_i x_{ij} - n_i p_i x_{ij} = -\sum_{i=1}^{N} n_i x_{ij}\frac{\partial}{\partial\beta_j}\left(\frac{e^{\sum_{j=0}^{k}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{k}\beta_j x_{ij}}}\right).$$ After using the general rules for differentiation;

$$\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_j} = -\sum_{i=1}^{N} n_i x_{ij} p_i(1-p_i)x_{ij} \quad \text{..... (3.11)}$$

Putting equation (3.11) equals to zero results to $k+1$ nonlinear equations with $k+1$ unknown variables. Solving a system of nonlinear equations is difficult, so that the solution must be numerically estimated by using an iterative process. The iterative solution need to be applied using Newton-Raphson method. To find the roots of equation (3.10) it is better to use matrix notation. The equation (3.10) can be written as $l'(\beta)$ and let $\beta^{(0)}$ represent a vector of initial approximations for each β_j, the initial step of Newton-Raphson can be expressed as

$\beta^{(1)} = \beta^{(0)} + \left[-l''(\beta^{(0)})\right]^{-1}.l'(\beta^{(0)})$ using matrix multiplication $l'(\beta) = X^T(y-\mu)$ Where μ is a column vector of length N with elements $u_i = n_i p_i$ and $l'(\beta)$ will be a column vector of length $k+1$ with elements $\frac{\partial l(\beta)}{\partial\beta_j}$. Also $l'(\beta) = -X^T WX$ where W is a square matrix of order N with diagonal elements $n_i p_i(1-n_i p_i)$ and zero everywhere else, then $l'(\beta)$, described using matrix multiplication as above, is a

$k+1 \text{ x } k+1$ square matrix with elements $\frac{\partial^2 l(\beta)}{\partial\beta_j \partial\beta_j}$ (Scott A., 2008). The initial step Newton-Raphson can be written as; $\beta^{(1)} = \beta^{(0)} + \left[X^T W X\right]^{-1} X^T (y-\mu)$.

This iteration will continue until there is no change between the elements of β from one to the next iteration. Then the maximum likelihood estimates will converge.

3.3 Variables Selection

The potential explanatory variables were examined to determine whether or not they are significant enough to be used in the model. Variable selection plays an important role in classification. When many variables are involved only that which is really required should be selected to eliminate the less significant variables from the analysis. Selecting a subset of the variables instead of the whole set of candidate variables is necessary due to the reasons that, it is cheaper to measure only a reduced set of variables, prediction accuracy may be improved through exclusion of redundant and irrelevant variables, the predictor to be built is usually simpler and potentially faster when fewer input variables are used and knowing which variables are relevant can give insight in to the nature of the prediction problem (Reunanen, 2003). Automatic backward elimination method was used since it is the simplest of all variable selection procedures, can be implemented without special software and it is objective since they use the p -value. Its procudere is as follows:

i. It starts with all the predictors in the model, the predictor with the highest p -value greator than α_{crit} is removed.
ii. Refit the model and go to (ii)
iii. Stop when all p -values are less than α_{crit}

IV. Results

4.1 Simulation Study Algorithm

A simulation of binary logistic data from Bernoulli distribution. In order to perform a simulation study, the parameters defined in the simulation setting section were used to generate a new random dataset in each iteration. A general logistic regression model is defined as: $\log it(\pi_i) = X_i\beta$ where, π is the probability of success for subject i, β is the vector of regression coefficients and X is the matrix of predictors possibly including an intercept.

4.2 Generating data and performing simulation

The steps followed in simulating a new dataset were as follows:

a) Generating the matrix of covariates X_i having N rows and $npred$ columns, with values sampled from a standard normal distribution,
b) Compute $X_i\beta$ by classical matrix multiplication, resulting in an $NX1$,
c) Calculate the 'true' probability of success π from the relationship,
$$\pi = \frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))} \quad \text{.....4.1}$$
d) Sample the vector of Bernoulli outcomes from an appropriate distribution using the 'true' probability of success π, for each observation in the data.
e) This final dataset is then used for subsequent analysis in that particular iteration $j, j = 1,2...Nsim$
f) Finally, data generation steps were repeated for the different parameter settings.

The Simulation function was executed for varying sample sizes and for a given sample size, several datasets were generated. Note that the regression coefficient estimation with $gls()$ and backward model selection algorithm were both run on the same dataset in each simulation. This allowed for conclusions to be based on exactly the same dataset in each iteration step.

4.3 Impact of automatic backward model selection criterion on the bias of the parameters for large sample size

Often, during model building, there is need to select a subset of possible predictors that best explain the variability in their observations. A more parsimonious model may be developed either through expert knowledge of the important predictors for the phenomenon under investigation, and/or using an automatic variable selection methods.

Automatic variable selection methods do not always result in meaningful models as they may exclude important predictors or result in meaningless models, for instance, in the case of dummy coded data. In order to investigate the behaviour of the backward variable selection algorithm in a logistic regression model, the researcher let the coefficient of x_4 in the simulation $\beta_4 = 0.0001$, signifying negligible contribution of x_4 in the model. Moreover, a specified wide range for the coefficients of the remaining predictors is used in order to assess the impact of the magnitude of a predictor when included in the model by the backward selection algorithm.

The backward selection methods involves starting by fitting a model with all the variables of interest, then the least significant variable is dropped. The process continues by successively re-fitting reduced models and applying the same rule until all remaining variables are statistically significant. The results in table 4.1 show the proportion (%) of times each predictor was included in the model.

Table 4. 1: Proportion (%) of times each predictor was included in the model

Sample Size	x1	x2	x3	x4	x5	x6
30	89.8	98.4	87.6	25.6	41.6	45.4
60	97.8	100.0	98.6	20.0	51.6	54.2
120	100.0	100.0	100.0	18.0	62.4	79.2
300	100.0	100.0	100.0	19.4	91.0	97.4
600	100.0	100.0	100.0	18.0	99.6	100.0

The result of backward selection indicates that; the relative magnitude of the true regression coefficients has an impact on the inclusion of the respective predictors in the model. And for small coefficients, the sample size has an impact on their inclusion in the model. This further emphasizes the need for adequate samples per predictor in the estimation of logistic regression.

4.6 Application of Logistic Regression to Diabetes Data

Summary statistics for continuous predictors

The summary statistics for the continuous predictors; Age, systolic pressure, diastolic pressure, BMI and blood sugar in terms of mean and standard deviation are as shown.

	Age	**Systolic Pressure**	**Diastolic Pressure**	**BMI**	**Blood Sugar**
Mean	44.64	120.23	75.22	25.15	**8.85**
SD	**19.06**	**23.70**	**12.03**	**8.90**	**8.34**

The mean age of the patients was 44.6 years with standard deviation of 19.06, Systolic pressure mean (120.23) and sd (23.7), diastolic pressure mean (75.22) and sd (12.03), BMI mean (25.15) and sd (8.9) and blood sugar mean (8.85) and sd (8.34).

4.6.2 Summary statistics for the categorical predictors is shown in the table below.

Table 4. 2: Summary statistics for continuous predictors

The researcher sought to find out the summary statistics for continuous predictors; diabetes status, gender, Visit type and blood sugar test type.

Diabetes	Gender	Visit	BloodSugar
No :529	Female:675	First :569	First : 40
Yes:680	Male :534	Referal:634	Random:1169
NA	NA	Deferal: 6	NA

On the status of diabetes, 680 patients were diabetic while 529 were non-diabetic, 675 were female and 534 male, 634 of the patients were referral's, 569 had visited for the first time while 6 were deferral. On the type of blood sugar test 1169 of the patients was random and 40 was taken before taking any meal.

4.6.3 Automatic Backward selection criterion

First, we fit a logistic regression model of the diabetes status regressed against $x_1 = age$, $x_2 =$ systolic blood pressure, $x_3 =$ diastolic blood pressure, $x_4 =$ BMI, $x_5 =$ gender, $x_6 =$ visit type and $x_7 =$ Blood sugar. The logistic regression equation is defined as, $\log it(\pi_i) = \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_7 x_7$.

Table 4. 3: Logistic regression coefficients for the full model

Term	estimate	std.error	statistic	p.value
Age	-0.0291860	0.0066706	-4.375292	0.0000121
SexFemale	-15.3702239	1.3614003	-11.290010	0.0000000
SexMale	-15.6714024	1.3726764	-11.416677	0.0000000
SystolicPressure	0.1045439	0.0105808	9.880496	0.0000000
DiastolicPressure	-0.0446358	0.0130208	-3.428048	0.0006079
BMI1	0.1462542	0.0252389	5.794783	0.0000000
Bloodsugar	0.8361114	0.0685115	12.203955	0.0000000
BloodSugarRFRandom	-0.8530984	0.6320141	-1.349809	0.1770773

The overall test of significance for the regression coefficients is shown on table 4.5 above. Using the p values obtained the values that had significant values were considered to be taken in to the logistic regression model. All the predictors except the indicator of whether the blood sugar test was the first or random, were found to be statistically significant in explaining the diabetes status of patients. The test was carried out at $\alpha = 0.05$ level of significance.

4.6.4 Logistic regression coefficients for the full model.

Table 4. 4: Logistic regression coefficients for the full model.

	LR Chisq	Df	Pr(>Chisq)
Age	20.768710	1	0.0000052
Sex	166.054416	2	0.0000000
SystolicPressure	143.161270	1	0.0000000
DiastolicPressure	12.094450	1	0.0005057
BMI1	37.994524	1	0.0000000
Bloodsugar	533.274342	1	0.0000000
BloodSugarRF	1.806874	1	0.1788837

4.6.5 Reduced logistic regression model

The backward variable selection algorithm was further applied to the case study data in order to evaluate whether a more parsimonious model could be derived. Table 4.7 presents the logistic regression coefficients for the reduced model.

4.6.6 Reduced model: logistic regression coefficients for the reduced model

Table 4. 5: Reduced model: logistic regression coefficients for the reduced model

term	estimate	std.error	statistic	p.value
Age	-0.0293310	0.0066502	-4.410568	0.0000103
SexFemale	-16.2642404	1.2090287	-13.452319	0.0000000
SexMale	-16.5701606	1.2191539	-13.591525	0.0000000
SystolicPressure	0.1047026	0.0105343	9.939249	0.0000000
DiastolicPressure	-0.0437381	0.0129533	-3.376607	0.0007339
BMI1	0.1463122	0.0251046	5.828095	0.0000000
Bloodsugar	0.8334789	0.0682429	12.213417	0.0000000

The final model from the automatic selection algorithm dropped the blood sugar test type which was found not to be statistically insignificant in the full model. There was no much disparity in the regression coefficients between the full and the reduced model.

4.6.6 Logistic regression coefficients for the reduced model

The Wald statistic was used to assess the contribution of individual predictors or the significance of individual coefficients in a given model. Table 4.8 presents the statistical significance of individual regression coefficients (βs) tested using the Wald Chi-square statistic.

Table 4. 6: Logistic regression coefficients for the reduced model

	LR Chisq	Df	Pr(>Chisq)
Age	21.09510	1	0.0000044
Sex	401.99232	2	0.0000000
SystolicPressure	144.14423	1	0.0000000
DiastolicPressure	11.73598	1	0.0006130
BMI1	38.48045	1	0.0000000
Bloodsugar	531.99523	1	0.0000000

Table 4.8 presents the regression coefficients for the reduced model using Wald Chi-square statistics. Age, Sex, Systolic pressure, Diastolic pressure, bmi and Blood sugar were significant predictors of diabetic status ($p < 0.05$) .

V. Conclusion

We can therefore conclude that; there is need to evaluate whether a more parsimonious model could be derived by selecting a subset of possible predictors that best explain the variability in their observations. Variable selection is a means to an end and not an end to itself. Variable selection helps to construct a model that predicts best or explains the relationships in the data. Forward selection criterion has drawbacks, including the fact that each addition of a new variable may render one or more of the already included variables non-significant. Step wise methods use a restricted search through the space of potential models and use a dubious hypothesis testing based method for choosing between models. Other methods like AIC and BIC need to be combined with the ANOVA table for objective results. An alternate approach used in this research which avoids this shortcoming is backward selection. Backward elimination methods are faster, simpler and can be easily implemented without special software. They are much objective since they use only the p -value to determine which variables to keep or remove. Finally, on the application of the diabetes data the predictors used to fit the full logistic regression model; age, systolic blood pressure, diastolic blood pressure, BMI, gender, visit type and Blood sugar except the indicator of whether the blood sugar test was the first or random, were found to be statistically significant in explaining the diabetes status of patients. The test was carried out at $\alpha = 0.05$ level of significance. The Wald test statistic used to assess the contribution of individual predictors or the significance of individual coefficients for the reduced model showed that all the included predictors in the model were significant.

References

[1] Abdelrahman A.l., 2010. Applying Logistic Regression Model to the Second Primary Cancer Data. Ain Shams University. Egypt.
[2] Agresti, A. (2007). An introduction to Categorical Data Analysis (2nd ed). Wiley-Interscience.
[3] Al-Ghamdi A.S (2002). Using Logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention.
[4] Al-Ghamdi A.S., 2001 Using Logistic regression to estimate the influence of accident factors on accident severity
[5] Allison, Paul D. (2014) "Measures of fit for logistic regression." Paper 1485-2014 presented at the SAS Global Forum, Washington, DC.
[6] Bewick, V. Cheek L & Ball, J. (2005). Statistics review 14: Logistic regression. Critical Care, London England.
[7] Burns N. and Grove, S. (2011). The practice of Nursing Research: Appraisal, Synthesis and Generation of Evidence. Maryland Heights Missouri: Saunders Elsevier.
[8] Cabrera, A. (2013). Logistic regression analysis in higher education: An applied perspective. Higher education: Hand book of Theory and Research Vol10.
[9] Chenge. (2010). Risk factors for type 2 diabetes mellitus among patients attending a rural Kenyan hospital. African journal of Primary Health Care and Family Medicine.
[10] Creswell, J. W. (2011). Educational research: Planning, conducting and evaluating quantitative and qualitative research. Merrill Prentice-Hall: Upper Saddle River, NJ.
[11] David G. Kleinbaum, M. K. (2010). Logistic regression. New York: Springer Science ,Business Media.
[12] Giancristofaro R.A & Salmaso L, (2003). Model Performance Analysis and model validation in Logistic regression. Statistica, 63(2).
[13] Harrell, F,E & Mark, D.B (1996) Multivariate Prognostic models: Issues in developing models; evaluating assumptions and ade-quacy, and measuring and reducing errors, Statistics in Medicine.
[14] Harrell, F. J. (2015). Tutorial in biostatistics; multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine.
[15] Hosmer DW, L. S. (2013). Goodness of fit tests for the multiple logistic regression model. NewYork: Communications in Statistics.
[16] Hosmer DW, Lemeshow SL (2000). Applied Logistic Regression. 2nd ed. Hoboken , NJ: Wiley-Interscience.
[17] Hosmer, D.W and Lemeshow, S. (2000) Applied Logistic Regression, John Wiley & sons Inc., New York
[18] Menard, S. (1995) Applied Logistic Regression Analysis, Thousand Oaks: Sage Publications
[19] Menard, S. (2010). Applied logistic regression analysis. Sage University paper Series on Quantitative Applicationsin the Social Sciences. Ca; Sage: Thousand Oaks.
[20] Park, Hyeoun-Ae (2013), An introduction to Logistic regression: From basic concepts to interpretation with particular attention to Nursing domain.
[21] Pencina MJ, D' Agostino RB Sr, D' Agostino Jr, Vasan RS (2008) . Evaluating the added predictive ability of a new marker: from area under ROC curve to reclassification and beyond. Stat Med.
[22] Peng C.Y, T. S. (2013). The use and interpretation of logistic regression in higher education journals. Research in Higher Education.
[23] Peng, C.-Y. L. (2013). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research .
[24] Prempeh E.A., (2009). Comparative Study of the Logistic Regression Analysis and the Discriminant Analysis. University of Cape Coast.
[25] Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research.
[26] Royston P, Altman DG (2010), Visualising and assessing discrimination in the logistic regression model stat Med.
[27] Royston. (2014). The use of customs and other techniques in modelling continuous covariates in logistic regression. Statistics in Medicine.

[28] Steyerberg EW, V. A. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology.
[29] WHO. (2015). Prevention of diabetes mellitus, Technical report Series No 844. Geneva: World Health Organization.
[30] World Scientific, pp. 1–2
[31] Wright, R. (2008). Logistic regression. In L.G. Grimm& P.R. eds., reading and understanding multivariate statistics. Washington, DC: APA, A widely used recent treatment.
[32] Yan, Xin (2009), Linear Regression Analysis: Theory and Computing.

Solomon Mulei Mutava. "Impact of Automatic Backward Model Selection Criterion on the Bias of the Parameters in Logistic Regression Model." IOSR Journal of Mathematics (IOSR-JM) 13.4 (2017): 25-33.