MODEL SELECTION TECHNIQUES WITH APPLICATION TO PREDICTING THE DELIVERY WEEK OF A CURRENT PREGNANCY

Ukwajunor E.E and Akarawak E.E.E
Department of Mathematics, University of Lagos, Lagos, Nigeria

ABSTRACT

Model selection, sometimes, referred to as variable selection, is the process of selecting a subset of independent variables for use in model building. Variable selection arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is an uncertainty about which subset to use. Identifying the best subset among many variables is one of the hardest parts of model building process. The purpose of this article is to understudy some of the widely used variable selection techniques, and to apply these techniques to predicting the delivery week of a current pregnancy based on some anthropometric parameters. Random sample of two hundred expectant mothers were available and their anthropometric parameters of interest were extracted from their record in the gynaecological section of Lagos University Teaching Hospital (LUTH). The findings reveal that the "best" model for predicting the delivery week of a current pregnancy is the model with the three predictors: Height of the expectant mother, Cervical length and Human Chorionic Gonadotropin (HCG).

Keywords: Variable selection, Anthropometric parameters, Predictors, Model building.

INTRODUCTION

Building statistical model of response as a function of multiple explanatory variables is an important part of any statistical analysis and a common practice in various professions. However, a fundamental difficulty in statistical modeling is the choice of an appropriate model. This problem arises when a statistical model contain many parameters and one is confronted with the choice of selecting which measures are important in predicting the outcome variable. For example, suppose Y *is* the response variable and $x1, x2 \ldots, x_p$ is a set of potential explanatory variables which are made up of vectors of n observations. The problem of variable selection arises when one want to model the relationship between Y *and* a subset of $x1, x2\ldots, x_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when parameter is large and $x1$, $x2$ …,x_p is thought to contain many redundant variables. Leaving out important covariates introduces bias into the parameter estimates, while including unimportant variable weakens the prediction capability of the model.

Model selection procedures have been developed to address the situation. Variable selection and model- building techniques are used to identify the best subset of predictors to be include in a regression model. The procedures identify a small group of regression models that are "good" according to a specified criterion. A detailed examination can be made of a number of the more promising or "candidate" models, leading to the selection of the final regression model to be employed.

There are two approaches to the selection of independent variables. The first approach considers all possible regression models that can be developed from the pool of potential independent variables and identifies subsets of the independent variables which are "good" according to a criterion specified by the investigator. The second approach employs automatic search procedures to arrive at a single subset of the independent variable. Once a few subsets have been identified as "good" ones, a final choice of the model must be made.

This choice is aided by residual analyses and examinations of the influential observations for each of the competing models. Information from this analysis, together with prior knowledge of the phenomenon study, will be helpful in choosing the final regression model to be employed (Michael *et al., 2005)*

LITERATURE REVIEW

A number of studies have attempted to developed criteria for selecting a statistical model from set of candidate models, given data. The earliest developments of such selection criteria were based on attempts to minimize the mean squared error prediction. The most familiar of these criteria is the Mallows C_p. Mallow's (1973) recommended using C_p plots to help gauge subset selection (see also Mallow, 1995). Two of the other most popular criteria, motivated from different viewpoints, are the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). Edward (2008), showed that if $\hat{i}r$ *denote* the maximum log-likelihood of the Y_{th} model, AIC select the model that maximizes ($\hat{i}r - q_r$), whereas BIC selects the model that maximizes ($\hat{i}r - (logn)q_r/2$).

Min and Guogiang (2000) examined model selection criteria for neural network time series forecasting. The authors showed that BIC imposes greater penalty for model complexity than AIC. Hence the use of BIC for model selection results in a model whose number of parameters is no greater than that chosen by AIC. BIC gives a consistent estimate of the order of AR model than AIC. Thus in a real life application, BIC is often preferred to AIC since it is more reliable criterion for model selection.

This paper provides an overview of some variable selection techniques with application to predicting the delivery week of a current pregnancy. The use of variables to predict the delivery week of current pregnancy has got the attention of many researchers in recent years. For example, Ramanathan *et al.* (2003) examined the potential value of routine measurement of cervical length in singleton low-risk pregnancies at 37 weeks of gestation in the prediction of onset and outcome of labor. The result from the study revealed that measurement of cervical at 37 weeks can define the likelihood of spontaneous delivery before 40 weeks and 10 days.

Wozniak *et al.* (2014) estimated the potential value of elastographic evaluation internal cervical OS stiffness at 18 - 22 weeks of pregnancy in low risk, asymptomatic women in the prediction of spontaneous preterm delivery. Mariorosaria *et al.* (2015) determine the relationship between cervical dilation and time of delivery in women with preterm labor. The finding showed that dilation of the cervix and gestation age at admission is associated with the time interval of delivery in women with preterm labor.

This paper contributes to several existing work in this area. It understudies some variable selection techniques and applies these techniques to predicting the delivery weeks of a current pregnancy using relevant data from Nigeria.

METHODOLOGY

This study relies on data from the qynecological section of Lagos University Teaching Hospital. Random sample of two hundred expectant mothers were available and their anthropometric parameters of interest were extracted from their records. These parameters includes; the delivery week (Y), age of expectant mother ($X1$), weight of expectant mother ($X2$), height of expectant mother ($X3$), foetus age ($X4$), cervical length ($X5$), and cervical width ($X6$) of expectant mother and human chorionic gonadotropin ($X7$). These variables constitute pool of potential explanatory variables for a predictive regression model. The response variable is delivery week Y. A first-order multiple liner regression model based on all the predictor variables was fitted to serve as a starting point. The full model is given by model (1)

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \beta 6X6 + \beta 7X7 \qquad (1)$$

The next stage in the model-building process is to examine whether all these potential predictor variables in model (1) are needed or whether a subset of them is adequate. Subset of model (1) was obtained using R statistical software with the "leaps function" and the result is shown in Table 1. From the possible model identify in model (1), we need to determine one predictor model that do the "best" at meeting some well-defined criteria.

Criteria for Model Selection

There are good numbers of criteria for comparing the various regression models in all-possible-regression selection procedure. In this section, we consider five of the most commonly used measures, namely; R-square ($R2$), Adjusted R-Square ($R2$), Mallow's C_p statistic (C_p), Akaike Information Criterion (AIC_p), Bayesian Criterion (SBC_p

R-square (R^2_p): R-square is a measure of the proportion of variability in the data set that is accounted for by a regression model. It assumes that every independent variable in the model helps to explain variation in the dependent variable (Y) and thus gives the percentage of explained variation if all independent variables in the model affect the dependent variables (Iwundu and Efezino, 2015). The (R^2) criterion is given by the
Statistic (2)

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad (2)$$

Where;
SSR is the regression sum of squares,
SSTO is the total sum of squares
SSE is the error sum of squares
R-Square value increases as more variables are added to the model, thus it make no sense to define the "best" model as the model with the largest R-Square value. However, the R- square statistic can be used to find the point where adding more predictors is not worthwhile, because it yields a very small increase in R-Square value.

Adjusted R-Square (R^2_{adj}) : The R^2_{adj} tells us the percentage of variation explained by only those independent variables that truly affect the dependent variable (Y) and penalizes for adding independent variable(s) that do not belong to the model (Iwundu and Efezino, 2015). These selection procedures do not take account of the number of parameters in the model. The adjusted R – square is defined by the statistic (3)

$$R^2_{adj} = 1 \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SSTO}\right) = 1 - \left(\frac{n-1}{SSTO}\right) MSE \tag{3}$$

Since $\underset{adj}{\max}(R2)$ can never decrease as p increases, the adjusted coefficient of the multiple determination adjusted R- square has been suggested as an alternative criterion. The best model according to the criterion is the model with the largest R – square value. Notice from equation (3) that the adjusted R- square value is a function of Mean Square Error (MSE) and the MSE given by

$$MSE = \frac{SSE}{n-p} \tag{4}$$

Quantifies how far away over predicted responses are from our observed responses. Naturally, we want this distance to be small. From (3), it is obvious that the adjusted R = square increases only if MSE decreases. Thus, the best regression model is the one with the smallest MSE value.

Mallow's C_p statistic: The C_p statistic is a criteria to asses fits when models with difference numbers of parameters are being compared. The criteria addresses the issue of over fitting, in which model selection statistics such as the residual sum of squares always get smaller as more variables are added to the model. Mallow's C_p is defined by the statistic

$$C_p = p + \frac{(MSE_p - \sigma^2)(n-p)}{\sigma^2} \tag{5}$$

Where MSE_p is the mean square error from fitting model containing subset of $p -1$ predictors. In using the C_p criterion, we seek to identify subsets of X variable for which (i) the C_p value is small and (ii) the C_p value is near p. Subsets with small C_p values have small total mean square error.

Akaike Information Criterion and Schwarz' Bayesian Criterion: Akaike Information Criterion (AIC_P) and Schwarz' Bayesian Criterion (SBC_p) are two popular selection criteria that penalize models for adding predictors. AIC and SBC act as guard against over fitting. The more parameter you fit to your model, a penalty is imposed. The criteria are defined as

$$AIC_p = nlnSSE_p – nlnn + 2p$$
$$SBC_p = nlnSSE_P –nlnn + [lnn] \qquad (6)$$

where n is the sample size. Since there are $2^{p-}1$ to consider among $p - 1$ potential variables, obtaining AIC_p or SBC_p for each model can become very tedious and time consuming. One way around this is to use the stepwise regression. The stepwise type procedures are based on three different strategies, namely, Forward Selection (FS), Backward Elimination (BE) and Stepwise Regression (SR).

Forward stepwise selection adds one variable at a time based on the lowest residual sum of squares until no more variables continue to lower the residual sum of squares. Backward stepwise regression starts with all variables in the model and removes variables one at a time. Stepwise regression is the modification of forward selection. The procedure involves the reevaluation of all regressors that previously entered into the model via their partial F-statistic. A regressor added at an earlier step may be consider redundant due to the relationship between it and regressors now in the model. The procedure requires two cut-off values F_{in} and F_{out}. If the partial F-statistic for a variable is less than F_{out}, that variable is dropped from the model.

RESULT AND DISCUSSION

Data analysis was performed using R statistical software. The summary of the results is shown in Table 1.

The best model according to R – square and adjusted R - Square is the model with the two predictors; height of expectant mother and human chorionic gonadotropic. Using the Mallow's C_p criterion, the model with the smallest C_p (C_p = 4.2047) is the model with the predictors height of the expectant mother and human chorionic gonadotropic.

The forward and backward regression identifies the best model as one which includes height of the expectant mother, cervical length and human chorionic gonadotropic. However, the stepwise regression picks the model with only age of the expectant mother and chorionic gonadotropic. The result for stepwise regression (SR) is shown on Table 2. The stepwise regression procedure eliminates the variable $X2$, $X3$, $X4$, $X5$, $X6$ leaving $X1$ and $X7$. However, one limitation of the stepwise regression search procedure is that it presumes there is a single "best" subset of X variables and seek to identity it. But in reality there is often no "best" subset. Hence, it was suggested that all possible regression models with similar number of X variables as in the stepwise regression solution should be fitted to study whether some other subsets of X variable might be better.

Thus, in line with this suggestion and since most existing literatures found cervical length to be a significant variable in predicting current pregnancy (Ramanathan *et al.* (2003), Wozniak *et al.* (2014)), couple with the principle that the number of predictors be kept to a limited number consisting of three to six "good" subset (Michael *et al.*, 2005), we decided to keep the three variables $X3$, $X5$, $X7$ in our final model. The model with the three predictor variables was fitted in R with the residual plot shown in Figure 1. The reduced model is given by model (7)

$$Y = \beta 0 + \beta 3X3 + \beta 5X5 + \beta 7X7 \qquad (7)$$

Table 1: R^2, R^2 , C_p, values for All Possible Regression Model (1)

Variables in Model	P	R_p^2	$R_{a,p}^2$	C_p	Variables in Model	P	R_p^2	$R_{a,p}^2$	C_p
X_1	1	0.000214	0.0028	99.6796	$X_4X_6X_7$	3	0.3423	0.3288	4.880
X_2	1	0.02815	0.0022	99.4802	$X_4X_5X_6X_7$	4	0.3420	0.3285	4.9559
X_3	1	0.0130	0.0.0080	96.4507	$X_3X_5X_6X_7$	4	0.3420	0.3285	4.9636
X_4	1	0.0561	0.0.0513	83.6992	$X_3X_4X_5X_7$	4	0.3397	0.3261	5.6486
X_5	1	0.00548	0.0499	84.1133	$X_1X_3X_5X_7$	4	0.3380	0.3244	6.1371
X_6	1	0.00047	0.0045	100.1739	$X_2X_3X_5X_7$	4	0.3359	0.3223	6.7677
X_7	1	0.3008	0.2972	11.1951	$X_1X_3X_4X_7$	4	0.3351	0.3214	7.0085
X_5X_7	2	0.4545	0.4331	660.933	$X_1X_5X_6X_7$	4	0.3349	0.3213	7.0614
X_3X_7	2	0.3081	0.3011	11.0103	$X_2X_4X_5X_7$	4	0.3338	0.3201	7.3868
X_1X_7	2	0.3037	0.2967	12.2977	$X_2X_5X_6X_7$	4	0.3311	0.3174	8.1810
X_6X_7	2	0.3030	0.32959	12.5246	$X_1X_2X_5X_7$	4	0.3498	0.3330	4.6542
X_4X_7	2	0.3028	0.2958	12.5640	$X_3X_4X_5X_6X_7$	5	0.3442	0.3273	6.2993
X_2X_7	2	0.3016	0.2945	12.9367	$X_1X_3X_5X_6X_7$	5	0.3438	0.3269	6.4192

X_4X_5	2	0.1213	0.1124	66.3616	$X_1X_4X_5X_6X_7$	5	0.3438	0.3269	6.4374
X_3X_5	2	0.0686	0.0591	81.9731	$X_1X_2X_3X_4X_7$	5	0.3428	0.3259	6.7165
X_3X_4	2	0.06512	0.0556	83.0183	$X_2X_4X_5X_6X_7$	5	0.3421	0.3251	6.9229
X_4X_6	2	0.06460	0.0551	83.1665	$X_2X_3X_5X_6X_7$	5	0.3421	0.3251	6.9322
X_3X_7	2	0.3378	0.7277	4.2047	$X_2X_3X_5X_6X_7$	5	0.3399	0.3229	7.5749
$X_3X_4X_7$	3	0.3347	0.3245	5.1272	$X_1X_2X_3X_5X_7$	5	0.3361	0.3190	8.6993
$X_4X_5X_7$	3	0.3336	0.3234	5.4619	$X_1X_2X_4X_5X_7$	5	0.3354	0.3182	8.9231
$X_5X_6X_7$	3	0.3310	0.3208	6.2200	$X_1X_2X_5X_6X_7$	5	0.3520	0.3318	6.0014
$X_1X_5X_7$	3	0.3298	0.3196	6.5615	$X_1X_3X_4X_5X_6X_7$	6	0.3498	0.3298	6.6534
$X_2X_5X_7$	3	0.3120	0.3015	11.8512	$X_1X_3X_4X_5X_6X_7$	6	0.3498	0.3298	6.6534
$X_1X_3X_7$	3	0.3106	0.3001	12.2626	$X_1X_3X_4X_5X_6X_7$	6	0.3498	0.3298	6.6534
$X_1X_6X_7$	3	0.3097	0.2991	12.5266	$X_1X_3X_4X_5X_6X_7$	6	0.3498	0.3298	6.6534
$X_3X_4X_7$	3	0.3081	0.2975	13.0009	$X_1X_3X_4X_5X_6X_7$	6	0.3498	0.3298	6.6534
$X_2X_3X_7$	3	0.3068	0.2962	13.3758	$X_1X_3X_4X_5X_6X_7X_7$	7	0.3498	0.3298	6.6534

Table 2: Stepwise regression output for model (1)

Initial model

$y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$

Final model

$y = X_1 + X_7$

Step	DF	Deviance Resid.		DF	Resid. Dev	AIC
1				92	299.9688	125.8308
2	$-X_6$	1	0.2782902	93	300.1871	123.9236
3	$-X_2$	1	0.3961992	94	300.5833	122.0555
4	$-X_5$	1	0.6693078	95	301.2526	126.2779
5	$-X_3$	1	1.1172117	96	302.3698	118.6481
6	$-X_4$	1	2.5102619	97	304.8801	117.4748

The residual vs. fitted plot do not look as there appear to be pattern in the dispersion which indicates that constant error variance is apparent. In addition, some departure from normality is suggested by the normal probability plot of the residuals [Figure 1]. Thus a weighted least square was employed as a remedial measure.

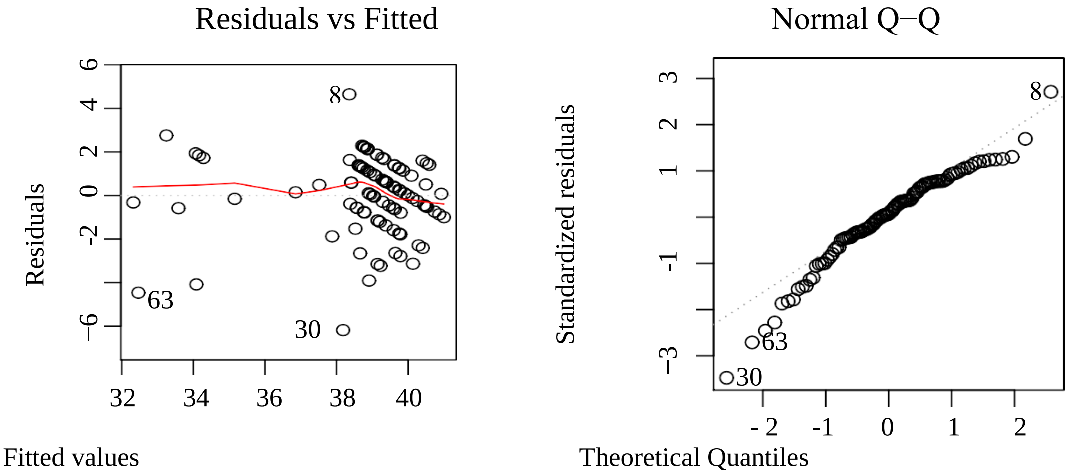

Figure 1: Residual plot the reduced model

CONCLUSION

The adequacy of variable selection techniques on model-building was examined based on some anthropometrics parameter of two hundred expectant mothers to predict their delivery week. Five selection techniques (*R* - square, *Adjusted R* - square, Mallow's C_p, AIC_p, SBC_p) were employed and all the five selection procedures identify the best model as one with the three predictor: height of the expectant mother, cervical length and the human chorionic gonadotropin. The model summary with the three predictor variables shows that height of the expectant mother is not significant. However, we decided to retain height of the expectant mother in the model following the principle that the number of predictors be kept to a limited number consisting of three to six ”good” subset.

The study did not consider interactions and quadratic terms in the model. Such terms have been found to be significant predictors; hence we suggest that the interaction and quadratic terms in model (1) be investigated. Based on our study, we recommends human chorionic gonadotropinic, cervical length and height of expectant mother as ”good” covariates in a regression model for predicting the delivery week of a current pregnancy.

REFERENCES

Edward, I.G (2000); *The Variable Selection Problem*, Journal of the American Statistical Association, 452.

Iwundu, M.P and Efezino, O.P (2015); *On the Adequacy of Variable Selection Techniques on Model Build- ing*, Asian Journal of Mathematics and Statistics, 8: 19 - 34, Available at doi: 10.3923/ajms.2015.19.34.

Mallows, C.L (1973); *Some Comments on C_p*, Technometrics 15, 661 - 676.

Mallows, C.L (1995); *Some Comments on C_p*, Technometrics 37, 362 - 372.

Mariarosaria, D.T, Voila, S and Tommaso, S (2015); *Relationship between cervical dilation and the time of delivery in women with preterm labour*, Journal of Research in Medical Sciences, 925 -929.

Micheal, H.K, Christopher, J.N, John, N and Willian, L.I (2005); *Applied Linear Statistical Model*, New York, McGraw Hill Companies, 417 - 428.

Min, Q and Guogiang, P.Z (2000); *An Investigation of Model Selection Criteria for Neural Network Time Series Forcasting*, Elsevier, 666 - 680.

Ramanathan, G, Yu, C, Osei, E and Nicolaide, K.H (2003); *Ultrasound examination at 37 weeks' gesta- tion in the prediction of pregnancy out: the value of cervical assessment*, Ultrasound in Obstetrics and Gyneacology, 598 - 603.

Wozniak, P, Piotr, C, Piotr, S, Pawel, M, Ewa, W and Tomasz, P (2014); *Elastography in predicting preterm delivery in asymtomatic low-risk women: a prospective observation study*, BMC pregnancy and childbirth, 238.