Multiple Linear Regression Analysis of Students' Academic Performance with Variable Selection

E. U. Oti[1]; R. A. Slink[2]; S. U. Enogwe[3]; R. Bekesuoyeibo[4]

[1,2.4]Department of Statistics,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria

[3]Department of Statistics,
Michael Okpara University of Agriculture, Umudike, Nigeria.
E-mail: eluchcollections@gmail.com[1]

Abstract — Regression analysis provides us with useful aid in the prediction of models; it examines and also explores relationships between variables. In this paper, the student's Grade Point Average (GPA) in Computer Science Department, Federal Polytechnic, Ekowe at the end of 2017/2018 first academic section which is the response (dependent) variable was regressed on four predictive (independent) variables namely: Chemistry, Mathematics, English, and Physics. The predictive variables are the Joint Admissions and Matriculation Board (JAMB) subjects offered by the respective students of the Computer Science department that were given provisional admission into the department. The forward, backward elimination, and stepwise selection procedures were used in selecting the best predictor variable subsets. It was observed from the results that the same subsets of predictor variables were selected for both forward and stepwise selection procedures. It is also seen that not all the linear functions of the unknown parameters $\beta_1, \dots, \beta_4$ are zero meaning that the explanatory variables have a significant influence on the dependent variable.

Keywords: *Multiple regression; Dependent variable; Predictor variables; Linear function; Sum of squares.*

I. Introduction

Multiple regression analysis is a statistical process for estimating the relationships between a dependent (response) variable Y and two or more independent (predictor) variables $X_1, X_2, \dots, X_k$ which is widely used for prediction and forecasting purposes. Notably, Bowerman and O'Connell (1997) added that we can more accurately describe, predict and control a dependent variable by using a regression model that employs more than one independent variable. The linear multiple regression models relating Y to $X_1, X_2, \dots, X_k$ is given as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \quad (1)$$

Then the deterministic component of the multiple regressions in Equation (1) will be

$$E(Y) = \mu_{y/X_1,X_2,\dots,X_k} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (2)$$

This is the mean value of the dependent variable Y when the values of the independent variables are $X_1, X_2, \dots, X_k$, where linearity implies that $E(Y)$ is a linear function of the unknown parameters $\beta_0, \beta_1, \dots, \beta_k$ that has to be estimated using sample data. ε is an error term that describes the effect on Y of all factors other than the values of the independent variables $X_1, X_2, \dots, X_k$.

The purpose of this paper is to carry out a multiple regression analysis of students' performance using variable selection procedures. The process of examining subset models and selecting one or more suitable prediction functions is often called the selection of variables (subset selection or subset analysis). The variable selection procedures that will be used are forward selection procedure, backward elimination procedure, and stepwise procedure rather than all-subset regression procedure considering the fact that it may not be the best function for predicting Y using $X_1, X_2, \dots, X_k$, but it may be an adequate prediction function for the problem (Graybill and Iyer, 1999), especially when there are k predictor variables in all, the number of possible subset prediction functions of the form $\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ is 2^k where $(X_1, X_2, \dots, X_n)$ is a subset of $(X_1, X_2, \dots, X_k)$.

The rest of the paper is organized as follows: Section 2 discusses the methodology of the variable

selection procedures. Furthermore, Section 3 describes the model selection criteria which are geared in selecting the best subset of predictions. While Section 4 shows the experimental results and discussion. Finally, Section 5 is the conclusion of the paper.

.

II. METHODOLOGY

The methodology of the variable selection procedures to be discussed is forward selection procedure, backward elimination procedure and stepwise selection procedure.

2.1 *Forward selection procedure*

This method starts with the simplest function β_0, and successively one variable is added to the model at a time in such a way that at each step the variable added is the best variable that can be added. For instance, describing the forward selection procedure (algorithm) using $k = 4$, that is, the total number of predictors under consideration is four which is denoted as $X_1, \dots, X_4$. At each step of the procedure, we will have a current model as the best candidate variable for adding to that model depending on any model selection criterion measure like Mean Square Error(s), Multiple Coefficient of Determination (R^2), Adjusted Multiple Coefficient of Determination ($\bar{R}^2$) and Mallow's C_p Criterion can be used, and each measure will select the same best candidate variable. Whether or not the variable added is actually added to the current model depends on whether a computed quantity denoted by F_C exceeds a criterion value which is denoted by F-in. The researcher chooses this criterion value to somewhat corresponding to a tabled F-Value with 1 degree of freedom in the numerator and $n-(p+1)$ degree of freedom in the denominator, where $(p+1$ stands for the number of β_1 in the model under consideration). This criterion value F-in is differently denoted in different statistical software packages. MINITAB uses the name **F to enter** to refer to the criterion value F-in. Since the function begins with β_0,

First step: for each predictor variable X_i, $(i = 1, \dots, 4)$ fit the model $\beta_0 + \beta_1 X_1$, using least squares we obtain $SSE(X_i)$, that is, the sum of squared errors using $\hat{\beta}_0 + \hat{\beta}_i X_i$ to predict Y. Choose the variable X_i that will result in the smallest value for $SSE(X_i)$ as the best candidate variable to be added the current model. Now $F_C = \frac{SSY - SSE(X_1)}{MSE(X_1)}$ where $MSE(X_1) = \frac{SSE(X_1)}{(n-2)}$ and SSY is the total sum of squares of the dependent variable SST. If $F_C \leq F\text{-}in$, then the algorithm stops and the original model B_0 is the final model. If $F_C > F\text{-}in$, then add X_i to the current model which makes it

$$\beta_0 + \beta_1 X_1 \qquad (3)$$

Second step: the current model is $\beta_0 + \beta_1 X_1$. The predictor variables that are not in this step are X_2, X_3 and X_4, for $i = 2,3,4$. At this point we fit the model $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ and obtain $SSE(X_1, X_i)$. Choose the variable X_i that result in the smallest value for $SSE(X_1, X_i)$ as the best variable to be added to the current model. Assume that the variable is X_2 then we calculate $F_C = \frac{SSE(X_1) - SS\ (X_1, X_2)}{MSE(X_1, X_2)}$ where $MSE(X_1, X_2) = SSE(X_1, X_2)/(n-3)$ like before, if $F_C \leq F\text{-}in$ the algorithm stops and we choose the model in Equation (3) as the final model, but if $F_C > F\text{-}in$ then add X_2 to the current model which makes it $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. This process continue until all the predictor variables are already included in the current model, which implies that there is no need to proceed further and the algorithm stops. In this case at the fourth step the final model becomes

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad (4)$$

2.2 *Backward elimination procedure*

This method begins with the present of a constant model β_0 with that which includes all of the available predictor variables, that is, $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ which proceed by successively eliminating one variable at a time from the model, such that in every step the variable removed is the variable contributing the least to the prediction of Y at that step. At each step of the algorithm, we will have a current model and will also label a predictor variable included in the current model as the best variable for deletion from the model. Whether or not this variable is deleted from the current model depends on whether quantity computed which is denoted by F_C is smaller than a criterion value that we call $F\text{-}out$. The model begins with

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_4 X_4 \qquad (5)$$

as the current model. The model is fitted using the least square method and calculate $SSE(X_1, \dots, X_4)$ using $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_4 X_4$ to predict Y.

First Step: X_1, X_2, X_3 and X_4 are available in the model of this step. For each predictor variable $X_i, i = 1, \dots, 4$, fit the model obtained by deleting this prediction variable from the current model and calculate the corresponding SSE leading us to consider the following from the models because $k = 4$.

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ *Entails that* X_4 *is omitted*
$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$ *Entails that* X_3 *is omitted*
$\beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$ *Entails that* X_2 *is omitted*
$\beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ *Entails that* X_1 *is omitted*

and the corresponding sums of squares error are $SSE(X_1, X_2, X_3)$, $SSE(X_1, X_2, X_4)$, $SSE(X_1, X_3, X_4)$, $SSE(X_2, X_3, X_4)$ respectively, suppose the first SSE is the smallest amongst them, that is, $SSE(X_1, X_2, X_3)$ it implies that if we want to delete one of the predictor variable in the current model, the best choice to variable to delete will be

X_4 because the three remaining predictors X_1, X_2 and X_3 are the best predictor subset models of the current model. The computed quantity becomes

$$F_C = \frac{SSE(X_1,X_2,X_3) - SSE(X_1,X_2,X_3,X_4)}{MSE(X_1,X_2,X_3,X_4)} \quad (6)$$

where $MSE(X_1, \dots, X_4) = SSE(X_1, \dots, X_4)/(n-5)$, if $F_C > F\text{-}out$, then the algorithm stops and the model in Equation (5) is chosen as the final model, which means that no variables are deleted in the first step and the variable in the model are $X_1, \dots, X_4$, but if $F_C \leq F\text{-}out$, then X_4 is deleted from the current model. When X_4 is deleted from the first step, then

$$\beta_0, + \beta_1 X_1 + \cdots + \beta_3 \quad (7)$$

is the remaining model containing variables X_1, X_2, X_3.

Second step: since Equation (7) is the current model for each predictor $X_i, i = 1,2,3$, fit the model obtained by deleting the predictor variable from the current model and calculates the corresponding SSE that leads us to consider the following three models.

$\beta_0, + \beta_1 X_1 + \beta_2 X_2$ $\quad Entails\ that\ X_3\ is\ omitted$
$\beta_0 + \beta_1 X_1 + \beta_3 X_3$ $\quad Entails\ that\ X_2\ is\ omitted$
$\beta_0 + \beta_2 X_2 + \beta_3 X_3$ $\quad Entails\ that\ X_1\ is\ omitted$

The corresponding SSE is $SSE(X_1, X_2)$, $SSE(X_1, X_3)$, $SSE(X_2, X_3)$ respectively. Suppose the smallest amongst these SSE is $SSE(X_1, X_2)$. We calculate

$$F_C = \frac{SSE(X_1,X_2) - SSE(X_1,X_2,X_3)}{MSE(X_1,X_2,X_3)} \quad (8)$$

where $MSE(X_1, X_2, X_3) = SSE(X_1, X_2, X_3)/(n-4)$ if $F_C > F\text{-}out$, then the algorithm stops and Equation (7) is chosen as the final model. Otherwise, if $F_C \leq F\text{-}out$, then delete X_i from the current model. When X_i is deleted in the second step, then

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (9)$$

Is the remaining model containing variables X_1 and X_2. We continue till the fourth step where X_1 is deleted and the procedure terminates.

2.3 *Stepwise selection procedure*

The stepwise selection procedure is the combination of the forward selection and the backward elimination procedure which allows re-examination at every step. Although, there are many versions of stepwise procedures but we will only discuss one in detail. The predictor variable is $k = 4$, then we start with an initial (current) model with no predictors β_0 or the full model $\beta_0 + \beta_1 X_1 + \cdots + \beta_4 X_4$ or any other subset model. The algorithm will proceed in two stages:

First stage, we start with the current model and perform the backward elimination procedure as many times as is necessary until no more variables can be deleted. If the current model is β_0, omit this stage and go to the second stage.

Second stage, we begin with the final model of the first stage and perform the forward selection procedure once, if a predictor variable is added to the current model, then go back to the first stage. If no predictor variable is added to the current model at this stage, then the procedure terminates because no variable can be added to the current model and no variable can be removed from the current model, in this case that current model is selected as the final model.

III. Model Selection Criteria

Many selection criteria for choosing the best have been proposed. These criteria are based on the principle of parsimony which suggests selecting a model with small residual sum of squares with as few parameters as possible. Hockings (1976) reviewed eight model selection criteria while Bendel and Afifi (1977) compared also eight criteria but not all the same as Hockings. A selection criterion is an index that can be computed for each candidate model and used to compare models (Kleinbaum et al. 1987). We shall consider four criteria: $s, R^2, \bar{R}^2, C_P$.

3.1 Mean square error (s)

The mean square error of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. The MSE is a measure of the quantity of an estimator, it is always non-negative, and values closer to zero are better. The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator (how widely spread the estimates are from one data sample to another) and its bias (how far off the average estimated value is from the truth (Draper and Smith, 1987).

3.2 *Multiple coefficient of determination*

The multiple coefficient of determination R^2 is the proportion of the total sum of squares of the dependent variables explained by the independent variables in the model

$$R^2 = \frac{SSR}{SSY} = \frac{SSY - SSE}{SSY}$$
$$= 1 - \frac{SSE}{SSY} \quad (10)$$

The objective is to select a model that accounts for as much of the variation in Y. Observe that in the above Equation (10), SSY is the same as SST, the use of the R^2 criterion for models building requires a judgment as to whether the increase in R^2 from additional variables justifies the increased complexity of the model (Rawlings et al., 1998).

3.3 *Adjusted multiple coefficient of determination*

The adjusted R^2 denoted as $adj.R^2$ or $\bar{R}^2$ is a rescaling of R^2 by degree of freedom so that it involves a ratio of mean square rather than sum of squares

$$adj.R^2 \; or \; \bar{R}^2 = 1 - \frac{MSE}{MSY} = 1 - \frac{(1-R^2)(n-1)}{(n-p)}$$
$$= 1 - \frac{(1-R^2)(n-1)}{(n-k-1)} \qquad (11)$$

3.4 *Mallows' C_P criterion*

The C_P criterion was proposed by Mallows (1973) and it is denoted as

$$C_P = \frac{SSE_P}{S^2} + 2P - n = \frac{SSE_P}{S^2} + 2(K+1) - n \quad (12)$$

Here S^2 is an estimate of σ^2, n is the number of observation, SSE_P is the sum of squares error from the p variable subset model..

IV. Experimental results and discussion

The data used in this paper is a primary data collected from the department of statistics, school of applied science, Federal Polytechnic, Ekowe Bayelsa State; 2017/2018 Academic session which sample size is 50 and the predictive variables are the Joint AdmissionsS and Matriculation Board (JAMB) subjects offered by the respective students of the department of Computer Science that were given provisional admission into the polytechnic, while the response variable is the Grade Point Average (GPA) of the respective students in the department at the end of first academic section which comprises of first and second semester examinations. The statistical software used in analyzing this paper is Minitab 16.0 version.

4.1 *Regression Analysis: GPA versus Chem, Maths, English, Physics*

The regression equation is

GPA = - 1.19 + 0.0196 Chem + 0.0120 Maths + 0.0147 English + 0.0227 Physics

Predictor	Coef	SE Coef	T	P
Constant	-1.1857	0.4678	-2.53	0.015
Chem	0.019553	0.007471	2.62	0.012
Maths	0.011991	0.007923	1.51	0.137
English	0.014707	0.008147	1.81	0.078
Physics	0.022677	0.007503	3.02	0.004

S = 0.285184 R-Sq = 65.7% R-Sq(adj) = 62.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	6.9975	1.7494	21.51	0.000
Residual Error	45	3.6598	0.0813		
Total	49	10.6573			

4.2 *Stepwise Regression: GPA versus Chem, Maths, English, Physics*

Forward selection. Alpha-to-Enter: 0.25

Response is GPA on 4 predictors, with N = 50

Step	1	2	3	4
Constant	0.1351	-0.3424	-0.9860	-1.1857
Physics	0.0444	0.0299	0.0225	0.0227
T-Value	6.81	4.21	2.96	3.02

P-Value	0.000	0.000	0.005	0.004
Chem		0.0238	0.0256	0.0196
T-Value		3.61	3.99	2.62
P-Value		0.001	0.000	0.012
English			0.0175	0.0147
T-Value			2.18	1.81
P-Value			0.034	0.078
Maths				0.0120
T-Value				1.51
P-Value				0.137
S	0.336	0.301	0.289	0.285
R-Sq	49.16	60.17	63.91	65.66
R-Sq(adj)	48.10	58.48	61.56	62.61
Mallows Cp	20.6	8.2	5.3	5.0

4.3 ***Stepwise Regression: GPA versus Chem, Maths, English, Physics***

Backward elimination. Alpha-to-Remove: 0.1
Response is GPA on 4 predictors, with N = 50

Step	1	2
Constant	-1.1857	-0.9860
Chem	0.0196	0.0256
T-Value	2.62	3.99
P-Value	0.012	0.000
Maths	0.0120	
T-Value	1.51	
P-Value	0.137	
English	0.0147	0.0175
T-Value	1.81	2.18
P-Value	0.078	0.034
Physics	0.0227	0.0225
T-Value	3.02	2.96
P-Value	0.004	0.005
S	0.285	0.289
R-Sq	65.66	63.91
R-Sq(adj)	62.61	61.56
Mallows Cp	5.0	5.3

4.4 ***Stepwise Regression: GPA versus Chem, Maths, English, Physics***

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is GPA on 4 predictors, with N = 50

Step	1	2	3	4
Constant	0.1351	-0.3424	-0.9860	-1.1857
Physics	0.0444	0.0299	0.0225	0.0227
T-Value	6.81	4.21	2.96	3.02
P-Value	0.000	0.000	0.005	0.004
Chem		0.0238	0.0256	0.0196
T-Value		3.61	3.99	2.62
P-Value		0.001	0.000	0.012
English			0.0175	0.0147
T-Value			2.18	1.81
P-Value			0.034	0.078
Maths				0.0120
T-Value				1.51
P-Value				0.137
S	0.336	0.301	0.289	0.285
R-Sq	49.16	60.17	63.91	65.66
R-Sq(adj)	48.10	58.48	61.56	62.61
Mallows Cp	20.6	8.2	5.3	5.0

Using Minitab statistical software in analyzing the above procedures, the same subsets of four independent (predictor) variables were selected for forward selection procedure and that of stepwise selection procedure. The $R^2 = 65.7$ in both the forward and stepwise selection procedure which shows the percentage of the GPA explained by the regression and also indicates how better the goodness of fit of the regression model to the sample data.

Now, considering the hypothesis testing of $H_0: \beta_1 = \cdots = \beta_4 = 0$ versus H_1: not all the $\beta_i's$ are zero. $F_{Cal.} = \frac{MSR}{MSE} = 21.51$ from the ANOVA table above, where $n = 50,\ k = 4,\ \alpha = 0.05,$ also the $F_{tab.} = F_{1-\alpha}; k, n - k - 1 = F_{1-0.05};\ 4,\ 50 - 4 - 1 = F_{0.95}; 4, 45 = 2.57.$
Since $F_{Cal.} > F_{tab.}$ We reject H_0 and accept H_1 (the alternative hypothesis) and conclude that not all the $\beta_i's$ are zero, meaning that the explanatory variables have significant influence on the dependent variable (GPA).

V. Conclusion

In this paper, we have discussed multiple regression analysis of students' performance using different variable selection procedure. From the experimental results, it was observed that forward selection procedure and stepwise selection procedure performed the same in terms of selecting the same variable subsets. The R^2 method is actually reasonable for the purpose of variable selection and it gives a clearer idea about the increase in variation explained by regression equation in terms of adding new variable in the model.

References

Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression: Journal of the American Statistical Association, 72, 46-53.

Bowerman, B. L. and O'Connell, R. T. (1997). Applied Statistics: Improving Business Process.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. Wiley and Sons, New York.

Graybill, A. F. and Iyer, H. K. (1994). Regression Analysis: Concepts and Applications.

Hockings, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics. 32 (1), 1-49.

Kleinbaum, G. D., Kapper, L. L. and Muller, E. K. (1987). Applied Regression Analysis and other Multivariate Method.

Mallow, C. L. (1973). Some comments on C_P. Technometrics 15, 661-675.

Rawlings, O. J., Pantula, G. S. and Dickey, A. D. (1998). Applied Regression Analysis: A Research Tool, Second Edition.

Appendix

Chem	Maths	Eng	Phy	GPA
41	56	64	55	2.88
54	57	53	49	2.54
61	64	52	48	3.12
57	61	56	58	2.45
48	62	54	57	3.08
72	64	65	68	3.42
57	55	48	61	2.55
55	54	62	54	2.24
54	52	50	46	1.89
56	62	49	54	2.44
45	52	54	50	2.18
62	56	48	62	2.56
53	57	52	60	2.66
56	62	56	58	3.00
60	47	44	62	2.75
61	54	58	63	2.86
52	42	46	48	1.90
48	56	53	51	2.06
42	50	62	54	2.34
65	60	63	67	3.34
53	52	50	54	2.26
56	60	48	58	2.45
49	52	54	52	2.24

Chem	Maths	Eng	Phy	GPA
54	56	62	54	2.64
64	62	60	56	3.06
66	64	48	56	3.04
70	72	54	62	3.46
72	66	54	62	3.56
65	64	58	60	2.88
44	48	52	56	2.34
53	57	55	53	2.33
48	50	53	45	1.90
51	54	53	62	2.54
48	50	54	56	2.67
54	55	62	48	2.48
54	58	52	50	2.86
58	56	64	72	3.11
48	46	52	54	2.58
52	56	64	67	2.65
56	62	66	70	3.26
68	74	62	82	3.68
48	46	51	54	3.00
56	52	50	61	2.44
52	62	46	50	1.87
48	45	49	53	2.24
53	63	56	58	2.86
34	52	50	45	1.98
54	62	62	56	2.32
56	59	56	54	3.03
62	49	62	65	3.34