MULTIPLE LINEAR REGRESSION MODEL: A STATISTICAL TOOL FOR PREDICTION OF SCORES OF FINAL YEAR MATHEMATICS DEGREE STUDENTS OF THE COLLEGE OF EDUCATION, AGBOR IN AFFILIATION WITH DELTA STATE UNIVERSITY, ABRAKA

Ohoriemu Blessing Okeoghene, Osemeke Reuben Friday
and **Aghamie Sunday Osiebuni**
Deperment of Mathematics, College of Education, Agbor
Email: oohoriemu@gmail.com

Abstract
Multiple linear regression is one of the most widely used statistical tools in diagnosing the performance of students in any examination. It is defined as a multivariate technique for assessing the correlation between a dependent variable(Y) and some combination of two or more independent variables(X1, X2, X3 . . Xp). In this paper, a multiple linear regression model comprises of three independent variables(X1, X2, X3) is developed to analyses the performance of final year mathematics Degree students of the College of Education, Agbor in affiliation with Delta State University, Abraka. The model is based on the data of student's scores in first tests, second test, and class attendance. The estimates both of the magnitude and statistical significance of relationships between the variables have been provided. Several statistical measures such as descriptive statistics, F calculated, T calculated, coefficient of determination(r^2), adjusted coefficient of determination, Mallows Cp Statistic, multicollinearity diagnostics and graphical residual plots, were used as a benchmark for selection of best subsets optimal regression models in a multiple regression diagnostics and a statistical tool for the analyzing the performance of final year mathematics Degree students of the College of Education, Agbor in affiliation with Delta State University, Abraka

1.0 Introduction

Multiple linear regression is used to analyzed the variability of a dependent variable(Y) using the information provided by the set of independent variables, X1, X2, X3, . . Xk as proclaimed by Pedhazur[1]. Regression the predictors against the dependent variable (Y) helps to determine and diagnose among predictors of interest which model is more statistical significant than the other.

Chatterjee, Price and Fox [2] considered varieties of numerical and graphical diagnostics for assessing the adequacy of regression model.

Multiple linear regression is defined as a multivariate technique for determining the relationship between a dependent variable(Y) and some combination of two or more predictor variables. (See, for example, Montgomery and Peck [3], Draper and Smith [4], Tamhane and Dunlop[5], and McClave and Sincich[6], among others, for details). It can be used to analyze data from causal-comparative, correlational, or experimental research. It can handle interval, ordinal, or categorical data.

Multiple linear regression is one of the most widely used statistical techniques in educational research. It is regarded as the "Mother of All Statistical Techniques." For example, many colleges and universities in Nigeria developed regression models for predicting the GPA of incoming fresh students. The predicted GPA can then be used to make admission decisions. In addition, many researchers have studied the use of multiple linear regressions in the field of educational research, business research, project analyses and data analyses in the areas of experimental design and students' performance.

Regression analysis has numerous importance and applications in every field such as engineering, chemical science, statistics, data analyses, economics, management, life biological sciences, and social sciences. Regression models are used for data description, parameter estimation, prediction, estimation and control. In this paper, a multiple linear regression model is developed to analyze the performance of the Mathematical scores of the final year Degree students of College of Education, Agbor in affiliation with Delta State University, Abraka.

1.1 Scope of Study

This research work is limited to three independent variables (X1, X2 and X3) to predict the dependent variable(Y). Many researchers have work with more than three independent variables in their analysis.

In addition, there are various regression model like linear regression model, log linear regression, quadratic regression model, cubic regression model and logistic regression model. The sample data was derived and was obtained from the scores of the final year mathematics Degree students of the College of Education, Agbor in affiliation with Delta State University, Abraka. The variables include Yi = Exam Performance, X1 = First Test Scores, X2 = Second Test Scores, X3 = Class Attendance

2.0 Methodology

2.1 Multicollinearity

Multicollinearity among the set of predictors is detected using Tolerance value, Variance Inflation Factor (VIF) and Correlation Matrix (CM).Tolerance value is defined as

$$1-r^2 \quad \textbf{(1)}$$

And VIF is defined as

$$\frac{1}{1-r^2} \quad \textbf{(2)}$$

Where r^2 is the coefficient of multiple determination of as regression of independent variables on all other independent variables.

2.1.1 Condition for Noticing Multicollinearity Problem

(i) Tolerance value less than 0.20 or 0.10.

(ii) Variance Inflation Factor (VIF) of 5-10 or above indicates problem. (David e tal [7].

(iii) For correlation matrix, two pair predictors must have a correlation close of [1 or -1] to indicate a collinearity problem. Brien,[7]. The simplest and most obvious means of identifying collinearity among two set of predictors is the examination of correlation matrix. Two predictors greater than 0.90 is an indication of substantial collinearity (David e tal [8].

2.2 Assumption of Regression that have to be validated to Ensure Good Regression Model

(i) The graphical plot of standardized residuals on the vertical axis against the set of predictors on the horizontal axis are used to evaluate the appropriateness. If the fitted model is appropriate for the data, there will be no apparent pattern in this plot of the residuals. The first assumption is residuals hovering around mean of zero in the scatter plots. This indicates normality validation.
(ii) The errors are all normally distributed showing evidence of linearity and constant variances of the error term. This is diagnosed using histogram of residuals, and Probability Plots. The normality plots give a graphical representation of the extent to which the data do not depart from normality, that is, the extent to which the little boxes in which the residuals cluster around the straight line. The Kolmogorov-Smirnov and Shapiro-Wilk tests can be used to test for the normality of the error term. However, if the Kolmogorov-Smirnov and Shapiro-Wilk tests are not significant, then the P value will be greater 0.05 .Hence, assumption of normality validation is met.
(iii) The constant variance of the error term (Homoscedasticity) can be evaluated using standardized residuals against the predictors. The normality of the residuals are evaluated using tallying the standardized residuals into a frequency distribution and displaying the results in a histogram. For Levene test, if P value < 0.05, there is a significance differences. This is an indication that variables are not homoscesdasticity in nature. But if the P value Test is greater than (>) 0.05, it means there are no significant differences. This is an indication that the variables are not homoscesdasticity. The variances are not equal.

2.3 Tests of Significance for Individual Parameters

The test statistic for T calculated for multiple variables is defined as

$$T = \frac{b_{i,i=1,2,3}}{sei} = \frac{Regression\ Coefficients}{standard\ error\ of\ estimates} \quad \textbf{(3)}$$

2.3.1 Critical Value for T calculated

$$T_{n-1}\alpha = T_{n-1}0.05 \quad \textbf{(4)}$$

2.3.2 Decision Rule

Reject H_0 if $|T| > T_{n-1}0.05$ **(5)**

Where

T = T Test statistics
N = Number of sample sizes
& = level of significance (0.05)

2.4: Testing for the Significance of Overall Selected Multiple Regression Models

After using residuals to ensure that the multiple regression models is appropriate, the next step is to determine whether there is a significant relationship between the dependent variable and the set of independent variables as given. The F test is used when there are multiple independent variables

2.4.1 Declaration of the Hypotheses of F Test

$H_0 = b_0 = b_1 = b_2 = b_3 = 0$ **(There is no linear relationship between the dependent variables(Y) and the set of selected independent variables (X1, X2, and X3)**

$H_1 = b_0 \neq b_1 \neq b_2 \neq b_3 \neq= 0$ **(There is a linear relationship between the dependent variables(Y) and the set of selected independent variables(X1, X2, and X3)**

However, if the null hypotheses are true and all slope coefficients are simultaneously equal to zero, the overall regression is not useful for prediction or descriptive purpose

Table 1: Sample ANOVA Table for Testing Significance of Regression Coefficients

Source	DF	SS	Mean square	F	F critical	Rejection region	Acceptance region
Regression	P	SSR	MSR = $\frac{SSR}{P}$	F = $\frac{MSR}{MSE}$	$f_{p-1,n-p-1}\alpha$	$f > f_{p-1,n-p-1}\alpha$	$f < f_{p-1,n-p-1}\alpha$
Error	N – P - 1	SSE	$MSE = \frac{SSE}{N-P-1}$				
Tota	n- 1	SST					

2.5: The Coefficient of multiple determination (r^2) and Adjusted r^2 Criterion

The coefficient of determination (r^2) is a statistic that gives some information about the goodness of fit of a model. In a regression model, r^2 is a statistical measure of how well the regression line approximates the real data points? Everett [9] defines r^2 as the coefficient of multiple determinations. r^2 Is the square of the correlation between two variables. Nagelkere [10] indicates that r^2 in the range of [-1 to 1] perfectly fits the regression line.

The coefficient of multiple determinations is equal to the regression sum of squares divided by the total sum of squares.

The model with the highest value of r^2 provides the closet fit. However, the major drawback of r^2 is that as the model increase, r^2 goes high whether the extra variables provides important information about the dependent variable(y) or not. Therefore, it makes no sense to define the best model with the largest r^2 value. A common way to avoid this problem is to use the adjusted version of r^2 instead of r^2 itself. The adjusted r^2 statistic for a model with P independent variables is given as

$$R^2_{adj} = 1 - \left[\left(1 - R^2_m\right) \frac{n-1}{n-p-1} \right] \quad \textbf{(6)}$$

Where

P = number of predictors in the regression equation, n = number of sample sizes, R_m^2 = coefficient of multiple determination. The r^2 adjusted does not necessarily increase when the number of predictors increase. It increases when the data has significant effects on the model. According to coefficient of determination and adjusted r^2 criterion, one should choose the model that has the largest r^2.

2.6 Mallows C_p Statistic

Gilmour [11] and Mallows [12] sees Cp Statistics as a measure for assessing the fit of regression model that has been estimated using ordinary least squares. It is applied in the context of model selection where a number of predictors available for predicting some outcome and the goal are to find the best model involving a subset of these predictors.

Mallows [12] have suggested using Cp as the best criterion for choosing a model among alternative competitive models. The model are unbiased when **Cp ≤ P +1**. For other illustrations and comment on interpretation sees Mallow [12], Goldman and Toman [13] or Daniel and Wood [14]. One disadvantage of Cp is that it seems to be necessary to evaluate Cp for all or most of the possible subsets to allow interpretation. The Cp statistic as defined by Mallow [12] is defined as

$$C_p = \frac{\left(1 - R_m^2\right)\left(n - T\right)}{1 - R_T^2} - \left[n - 2(p+1)\right] \qquad \textbf{(7)}$$

Where

P denotes the number of predictors included in the regression model

T denotes total number of parameters (including the intercept) to be estimated in the full regression model

R_m^2 Denotes coefficient of multiple determination for a full regression model that has P predictors

R_T^2 Denotes of multiple determination for a full regression model than contains all T estimated parameter

3.0: Data Analysis and Interpretation of Results

The R^2 value in the regression output indicates that only 53.3 % of the total variation of the Y values about their mean can be explained by the predictor variables used in the model. The adjusted R^2 indicates that only 49.3 % of the total variation of the Y values about their mean can be explained by the predictor variables used in the model. In real sense, the influence of the variables on Y is fairly average. It is an indication that the students perform fairly on exam performance. The standard error for full model is 13.13755

Table 2: Analysis of Variance Table

MODEL	SS	DF	MS	F	P	F Critical	Decision	Conclusion
Regression	6890.600	3	2296.867	13.308	0.000	3.23	Reject H_0	There is a linear relationship between the variables
Residual	6040.836	35	172.595					
Total	12931.43	38	2296.87					

Determination of the F Critical Value

$$f_{p-1,n-p-1}\alpha = f_{3-1,39-3-1}0.05 = f_{2,35}0.05 = 3.23 \quad \textbf{(8)}$$

From Table 2, the F calculated value is 13.308. The decision is that since the f calculated of 13.308 is greater the. F critical value of 3.23, we reject null hypotheses and conclude that there is a linear relationship between the dependent variable(Y) and at least some of the independent variables.In simple terms, at least one of the independent variables(X1, X2 and X3) is related to Y.

From Table 2, we observed that the P - Value is 0.000 which is less than 0.05 level of Significance.

This implies that the model estimated by the regression procedure is significant at an alpha level of 0.05.

Thus at least one of the regression coefficients is different from zero(0).

The P value <0.05 means the regression is significant, that is there is a differences in the given regression variable

Table 3: Unusual Observations of Data's from the Variables

Case No Outliers	ForOutlying Residuals	ZesClass Attendance No	Fitted Value		Zres Residuals
13	2.023	85.00	58.4223	26.57773r	2.11R
38	-2.103	45.00	72.6269	-27.62691r	-2.14R

R denotes an observation with a large standardized residual, which is an indication of Unusual Observations. Observations 13 and 38 are identified as unusual because the absolute value of the standardized residuals are greater than . ± 2. SPSS uses $\pm$3 for benchmark for detention of outliers. Hence observation 13 and 18 should be remove from the model and the model re-estimated for further analysis. Researchers should remove observation 13 and 18 from the model and re estimate the differences. Re-estimating the model without observation 13 and 38 gives the following output $R^2 = 0.627$, $Adjr^2 = 0,.593$, standard error = 11.703, P value = 0.000. The improve regression model will be

$$Y_i = 4.528 + 0.288_1 + 0.316 X_2 + 0.336 X_3 \qquad (9)$$

Examine and Testing the Adequacy of Multiple Regression Model for Predicting the Student's Final Exam Grade in a Mathematics Class Exam

Table 4: Computed values of Regression Coefficients, T Calculated, P value, Partial Correlation

RM	**Regression**	**T Calculated**	**P value**	**PARTIAL**
constant	8.978	.922	0.363	.275
X1	.247	1.694	0.099	.430
X2	.338	2.815	0.008	.393
X3	.290	2.530	0.016	.275

The P-values for the estimated coefficients of X_2 and X_3, are respectively 0.008 and 0.016 which is less than 0.05, indicating that they are significantly related to Y. This is an indication of a differences in the variables. The P-value for X1 is 0.099, indicating that it is probably not related to Y at an alpha-level of 0.05.

Decision rule for T Test

Reject Ho when the T Calculated is greater than t critical or accept Ho when t calculated is less than the t critical

$T_{n-1}0.05 = T_{39-1}0.05 = T_{38}0.05 = 1.6849$ or -1.6849. (10)

Variable X1 , X2 and X3 are accepted because its calculated T calculated as seen in Table 4 is greater than $\underset{-}{+}1.6849$. In conclusion, variable X1, X2 and X3 are good for prediction purposes.The model becomes

$$\hat{Y}_i = 8.98 + 0.247 X_1 + 0.338 X_2 + 0.290 X_3 \qquad (11)$$

Table 5: Testing For Multicollearity Using Correlation Matrix

	EXAM PERFORMANCE(Y)	**FIRST TEST SCORES(X1)**	**SECOND TEST SCORES(X2)**	**CLASS ATTENDANC**
EXAM PERFORMANCE(Y)	1.000	0.573	0.613	0.474
FIRST TEST SCORES (X1	0.573	1.000	0.581	0.306
SECOND TEST SCORES (X2)	0.613	0.581	1.000	0.227
CLASS ATTENDANCE (X3)	0.474	0.306	0.227	1.000

Table 6: Testing For Multicollinearity Using Tolerance Value And Variance Inflation Factor (VIF)

Statistics	**X1**	**X2**	**X3**
Tolerance Value	1.587	1.516	1.108
VIF	0.630	0.660	0.903

Table 6 has all the predictors of a VIF less than 2.The bivariate correlation matrix of Table 5 shows that no linear dependence exists among the predictors. Table 5 and 6 shows no evidence of multicollinerairyty problems among set of predictors

4.2 Best Subsets Regression:

Another important criterion for assessing the predictive ability of a multiple linear regression model is to examine the associated Mallows Cp statistic. The best subsets regression method is used to choose a subset of optimal predictor variables so that the corresponding fitted regression model optimizes the statistic.

Table 7: Best subsets regression output manually computed

Models	**Cp**	**R^2**	**ADJ R^2**	**P+1**	**P**	**Model Inclusion**
X1	15.29	0.329	0.311	2	1	No
X1,X2	8.445	0.447	0.417	3	2	No
X1, X2, X3	4	0.427	0.395	4	3	Yes ***
X1X3	9.944	0.533	0.493	3	2	No
X2	11.77	0.376	0.360	2	1	No
X2,X3	4.848	0.495	0.466	3	2	No
X3	23.08	0.225	0.204	2	1	No

The model with (***) satisfies the condition for $C_p \leq P+1$ and are good for model inclusion according to Mallows Cp statistic.

Using equation 7, we can compute Mallows Cp for the model containing X1, X2, X3

N = 39, P = 3, T = 4,P^2p = 0.533, R^2T = 0.533

$$C_p = \frac{\left(1-R_p^2\right)(n-T)}{1-R_T^2} - \left[n-2(p+1)\right] \qquad (12)$$

$$C_p = \frac{(1-0.533)(39-4)}{1-0.533} - \left[39-2(3+1)\right] = 4 \qquad (13)$$

The model X1, X2, X3 is hereby selected for further analysis and prediction. The fitted linear regression model is finally

$$\hat{Y}_i = 8.98 + 0.247X_1 + 0.338X_2 + 0.290X_3 \qquad (14)$$

From this model, the following can be concluded

(i) Holding constant the effect of X2 and X3, for each increase of one person in X1, we predict that the average Y will increase by 0.247.

(ii) Holding the X2 and X1 Constant, we conclude that the average value of X3 will increase by 0.290

(iii) Holding X1 and X3 constant we conclude that average value of X2 will increase by 0.338

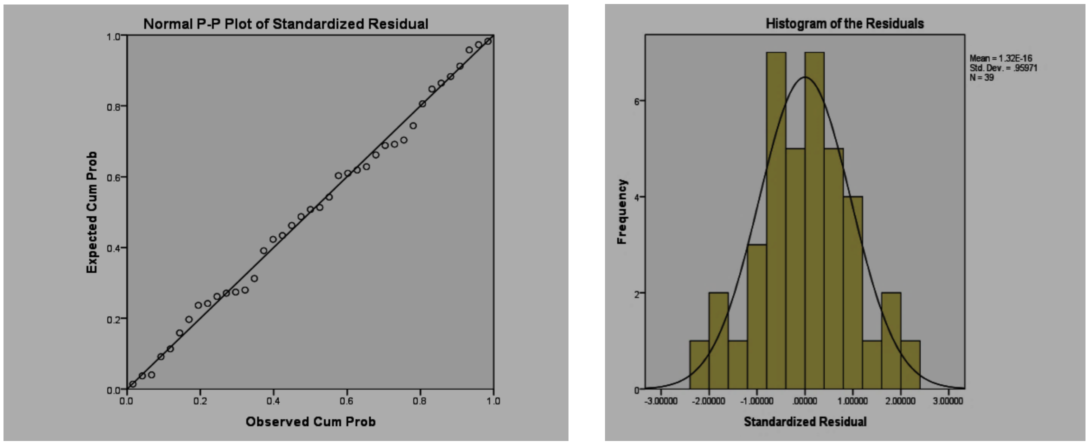


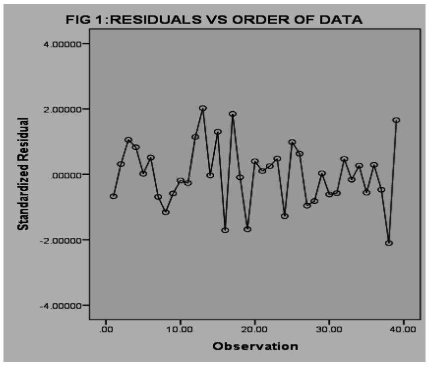
Fig 4.1: Graphical analysis of Probability Plots and Histogram of the residuals in checking for normality assumption and goodness of fit of regression model

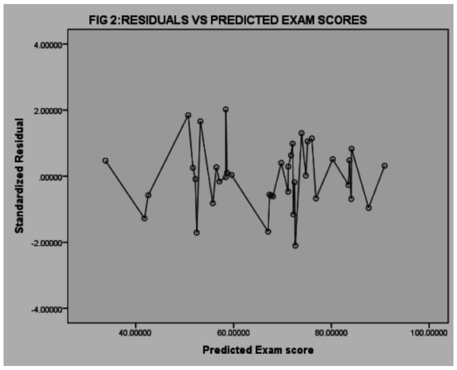
(i) From Normal probability plot, we observe that there exists an approximately linear pattern. This indicates the consistency of the data with a normal distribution. The outliers are indicated by the points in the upper-right corner of the plot.

(ii) The histogram of the residuals indicates that no outliers exist in the data. The histogram are evenly spread showing evidence of normality distribution.

5.0 Assessing Residuals scatter plots

The estimated final model will be used to through residuals to examine the regression cited assumption validation when there is minimal correlation among the predictor variables (Steven, 2019).





(i) Statistical Packages for Social Sciences (SPSS) were used to display this graph above. The graphical plots of standardized residuals against the predicted exam score or fitted value in Fig 2 provides an idea about the overall model equations. The plot gives an insight about the partial regression plots and the plots of standardized residuals against each of the independent variables. Validation of the regression assumption shows that the partial regression plots validates the regression assumption. Also, the scatter plots of Fig 2 shows that the graph follows a linearity pattern, and they are scatter around zero, showing evidence of constant variance of errors term. The finding indicates homoscedasticity in the multivariate of scatter plots of the residuals against set of independent variables and the scatter plots of standardized partial plots is an indication that the overall equation is linear

(ii) The plot for standardized residuals versus order is also provided in Fig 1 .This is an indication that the plot of all residuals in the order that the data was collected. It is used to find non-random error, especially of time-related effects.

Conclusion

The fitted multiple regression model for predicting the student final examination grade is given by

Yi = 8.98 + 0.247X1+ 0.338X2+0.290X3

From the above analysis, it appears that our fitted multiple regression model for predicting the student's final examination grade is useful and appropriate. In the presence of X1 and X3, X2 is a good predictor of Y. In the presence of X1 and X3, X2 is a good predictor of Y. In the presence of X1 and X2, X3 is a good predictor of Y.

As the values of R^2 and adjusted r^2 and are not very different, it appears that at least one of the predictor variables contributes information for the prediction of Y.

Also, since the test statistic value of F calculated from the data of 31.13 exceeds the critical value of 3.23, we reject the null hypotheses and conclude that the regression model is not significant, that is there is no differences in the variables. Hence, our multiple regression model for predicting the student's final examination scores seems to be useful and adequate, and the overall regression is statistically significant.

For future work, one can consider to develop and study similar models from the fields of education, social and behavioral sciences. One can also develop similar models by adding other variables, for example, the attitude, interest, prerequisite, gender, age, marital status, employment status, race and ethnicity of the student, as well as the squares, cubes, and, cross products of X1,X2 and X3. In addition, one could also study the effect of some data transformations on the variables.

References

Chatterjee, S Price, B. and Fox (1991). Useful and Recommended Practice. New York. Wiley & Sons Inc. New York.

Daniel, C. and F.S. Wood (1980). Fitting Equations to Data 2nd ed. New York:Wiley and Sons Inc New York.

David, K Hilderland and Lyman. O. (1991). Statistical Thinking For managers (3rd ed), PWS- Kent, publisher

David, K. Hilderland and Lyman, O. (1991). Statistical Thinking For managers (3rd ed), PWS- Kent, Publisher.

Draper, N. R., and Harry S. (1998). Applied Regression Analysis (3rd edition). New York: John Wiley & Sons, Inc.

Everett, B. S. (2002). The Cambridge Dictionary of Statistics (2nd Ed.). Cambridge University Press Publisher, New York.

Gilmour and Steven (1996). The interpretation of Mallows Cp- Statistics. Journal of the Royal Statistical Society, Seroies 45(1): pp. 45-56.

Goldman, J.W. and Torman, R.J. (1996). Selection of variables for fitting equations to data (2nd ed). Wiley & Sons Inc New York.

Mallows, C.I. (1973). Some Useful on Cp Technometrics 15,661-675.

McClave, J. T., and Sincich, T. (2006). Statistics (10th edition). Upper Saddle River, NJ: Pearson Prentice Hall.

Montgomery, D. C., and Peck, E. A. (1982). Introduction to Linear Regression Analysis. New York: John Wiley & Sons, Inc.

Nagelkere, N.J. (1991). A Note on a General Definition of the Coefficient of Determination. Biometrices; 78(3): pp. 691-692.

Pedhazur, E.J., 1997). Multiple Regression in Behavioral Research(3rd Ed). Orlando, FL:Harcount Brace.

Tamhane, A. C., and Dunlop, D. D. (2000). Statistics and Data Analysis: From Elementary to intermediate (1st edition). Upper Saddle River, NJ: Pearson Prentice Hall.