RESEARCH ARTICLE OPEN ACCESS

Predicting University Freshers' First-Year Performance from JME Score (A Case Study in ATBU, Bauchi)

Kefas Suwa Larson *, Dr.Nentawe Gurumdimma **
*(Department *of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria*
Email: larson_kefas@yahoo.com)
** (Department *of Computer Science, University of Jos, Plateau State, Nigeria*
Email :speakwithtawe@gmail.com)

--************************----------------------------------

Abstract:
This paper presents an attempt to predict students' academic performance (grades) in first-year based upon their scores in Joint Matriculation Examination (JME). In this study, simple linear regression is used, since two variables were majorly involved. Some sample data of JME scores and Gross Point Average (GPA) were used to build a model for predicting the mean values of the GPA when the values of JME Score are known, and to obtain the measure of errors involved in such prediction. The results show that the correlation coefficient, R is 0.137 which implies that the relationship between GPA and JME scores is very weak but positive. The coefficient of determination R^2 achieved 0.0188, meaning that our regression model can only explain 1.9% of the observed variability in the data. The other 98.1% is unexplained due to many other factors that can affect a student performance while in the University. Inference statistics about the fitted regression equation were computed, which gave the following results: (F-test =1.880 vs F-critical=3.938, t-test =1.371 vs t-critical =1.984). ANOVA results also show that F-value =1.880, with F-critical value =3.938, and p-value =0.173 at a significance level of 0.05. Considering the results of the different tests conducted, the null hypothesis cannot be rejected (p>0.05). Thus, we conclude that there is no strong relationship between JME score and GPA of University Freshers.

Keywords **—Academic Performance, University Freshers, GPA and JME scores, Linear Regression.**

--************************----------------------------------

I. INTRODUCTION

There are many statistical investigations in which the main objective is to determine whether a relationship exists between two or more variables. If such a relationship can be expressed by a mathematical function or formula, we will be able to use it for the purpose of making predictions. For instance, we may wish to estimate a man's future income, given that he has completed 4 years of university education. A training director may wish to study the relationship between the duration for new recruits' performance in a skilled job, and so on. In all these scenarios, the aim is the prediction of one variable from the knowledge of the other variable(s).

In this paper, we wish to predict students' academic performance (grades) in first-year based upon their scores in Joint Matriculation Examination (JME), otherwise known nowadays as Unified Tertiary University Examination (UTME). The study is specifically aimed at comparing the two variables, Grade Point Average (GPA) and JME Scores of University Freshers with the purpose of making prediction about their scholastic achievements in their first-year in the University and to estimate the amount of errors involved in such predictions.

This research wants to help the university community understands the impact of JME on the performance of students in higher institutions, and could form the basis for argument for Universities doing their entrance exams through other means such as post-University Tertiary Matriculation Examination (post-UTME), rather than only through the JME.

According to **[1]**, students with high intelligence tend to make high grades, but there are exceptions because of several factors other than intelligence that can affect school grades and achievement.

In **[2],** the authors observed that there are several works related to prediction of first-year students' academic performance in literature. But most of the models are complex and difficult to explain. Hence, a simple model that can be understood by even a lay man is needed for such predictions.

Other studies that have compared different techniques used to gain an understanding of which are the best predictors of students' academic performance while in the University include (**[9], [11], [13], [15], etc.**).

The authors in **[3],** stated that for over 200 years now, Regression Analysis is being employed to determine the relation between response variable (GPA) and explanatory variables (such as university entrance exams, achievement test, general aptitude test, etc.).

According to **[4]**, Linear Regression is the best prediction model to test the cause of one dependent variable's (e.g., first-year GPA) effect on one or more independent variables. This is because linear regression approach is so easy, with faster processing and for large datasets size.

In **[5],** the authors stated that the field of predicting students' performance is large and so far, the work has not converged sufficiently for researchers to be able to determine the state-of-art in this area. That, this solely depends on the methodology, data, usability and the application.

More so, most of the modern techniques that are used in predicting students' academic performance are built on the concept of statistical model of regression analysis (Data mining, machine learning, Deep learning, Artificial intelligence, Big Data Analytics, etc.).

In this study, we believe that a simple prediction models such as Linear Regression can perfectly be used to fit a model for the prediction of students' academic performance.

A. GPA and JME Scores

First, we note that JME is an examination set up by Joint Admission Matriculation Board (JAMB). It is one of the entry requirements for admission to first degree courses into Nigerian universities.

Usually, applicants who take JME may be admitted into part one (100-level) and are called Freshers. Freshers refers to students who have just started their first year at University. JME scores are the marks earned by candidates who take the JAMB's JME. Each University has its cut-off marks for a particular course which students have to reached before being admitted. The minimum entry requirements for admission into Nigerian University is 5 credit passes in English, Mathematics and 2 other subjects, depending on the department. For example, in Abubakar Tafawa Balewa University (ATBU), Bauchi, the general cut-off mark is 180, and cut-off mark by Faculty for the year 2019/2020 is as shown in Table I.

TABLE I
JAMB 2019/2020 CUT-OFF MARK FOR ATBU

Faculty/School	**JME Scores**	**Cut-off Score**
Agriculture & Agric Technology	English, Chemistry, Biology/Agric. and any one of Physics and Mathematics	220 Points
Engineering & Engr Technology	English, Physics, Chemistry and Mathematics	220 Points
Environmental Technology	English, Math and any of Physics, Chemistry, Biology, Geography, Fine Art, Biology or Economics	220 Points
Management Technology	2 Commercials from Economics, Commerce, Accounting, Business Method and any one of Chemistry, Physics, Biology or Agric. Science, Geography, ICT/Computer	190 Points
Science	English, Math and any 2 of Physics, Chemistry, Biology, Economics /Agriculture, Statistics, Further Math or Geography	190 Points
Education	Mathematics and any 2 of the following science subjects: Physics, Chemistry, Biology	190 Points
College of Medical Science	English, Biology, Chemistry and Physics	250 Points

Comparing the students' GPA and JME scores provide an excellent opportunity for University management to identify their level of previous knowledge, since it is usually assumed that students with strong secondary academic background demonstrate better prospects of performing well in University than those with low JME scores. To buttress this point, **[6]** stated that due to the notion that students with high scores in pre-university exams are most likely to succeed academically, universities and other tertiary institutions accord considerable importance to this topic. Students admitted with higher entry qualifications are generally expected to be well prepared for the university course content than those with lower qualifications.

B. The Aim and Objectives of the Study

The main aim of this study is to use some sample data of JME score and GPA to fit a model for predicting the expected values of GPA when the values of JME scores are known. There are three basic objectives of the study:

1. To provide estimates of the values of GPA from the values of the JME scores,
2. To obtain measures of the errors involved in this prediction, and
3. To find the measure of the degree of association or correlation between these two variables.

The rest of this paper is organized as follows: Section **II** describes Related Work. Section **III** presents the proposed Methodology. Section **IV** discusses the results of the implementation using MATLAB, Minitab and the simulation conducted, and finally, Section **V** concludes the study with recommendations, and highlights future research directions.

II. REVIEW OF RELATED LITERATURE

This section provides a review of the related work done by different researchers on students' entry exams scores compared with their academic performance in the University.

Few works such as **[8, 9, 12,13, 14, 16]** have studied the use of Multiple Linear Regression (MLR) to predict the academic performance of students, and found out that there was no significant relationship between university first-year GPA and entry exams (JME /UTME) scores. However, none of them was able to determine the actual reasons why UTME, which is the main national entrance examination into Nigerian Universities is not able to predict students' performance while in the University. The implication of this is that research on the relationship between the JME performance and University academic achievements of students is not yet conclusive.

The authors in **[7],** specifically focused their study on the problem of predicting the academic success of students in higher education. The study presented a clear set of guidelines to follow while using Educational Data Mining for success prediction. However, they observed that the problem with prediction using data mining, artificial neural networks, support vector machines, and Naïve Bayes algorithms is that they are considered to be black-box methods, being complex and uneasy to understand. Therefore, a simple technique like Simple Linear Regression (SLR) would be far better for the prediction of students' academic performance.

In **[3],**the authors aimed their paper at examining the factors affecting students' GPA at the college of science in Prince Sattam Bin Abdul-Aziz University (PSAU). The study focused on identifying the relationship between the students' GPA and the variables that affect it (e.g., secondary school level and achievement test). In this paper, two approaches of regression were applied; namely Ordinary Least Square Method (OLS) and Quantile Regression (QR) to investigate the relationship between GPA and secondary school rate, achievement test, gender, and department, to decide the independent variable impact power on the dependent variable. The OLS method used revealed a significant effect on all the four variables: secondary school rate, achievement test, department, and gender on GPA and department. However, the study has a limitation in terms of inadequate data in the predictor variables, such as, the students' attendance, study hours, place of residence, mother education, age, and teaching methods.

The research in **[10]** applied regression techniques on dataset of graduate and undergraduate students to find out if pre-university exams have real effects on the students' university GPA. They found out that though high school GPA affects the university GPA, pre-university exams like General Aptitude Test (GAT) and Achievement Test (AT) do not affect university GPA of students. However, their model used 6 independent variables (University GPA, High School GPA, GAT, AT, students' graduation year, students' enrolled year) and cannot make a perfect comparison with our research which is based on a single independent variable.

In **[11],** the authors used a predictive model called new Academic Performance Indicator (API) for the first term of first-year science degrees students at La Laguna University, Spain. They applied MLR model to identify which measurements of previous students' performance are good predictors of academic success in the first year of study. They found out that the best predictor of academic success is high school GPA (aggregate of JME and WAEC scores) and the GPA of the student in their first-year in university. However, their study is limited to (Mathematics, Chemistry, Physics, Biology and Environmental Sciences) and need to be extended to analyze other degrees to determine whether those findings can be generalized to all courses offered in University.

The authors in **[15],** compared admission exams versus high school scores and found out that secondary score is a stronger predictor of students' performance at the university. One of the main conclusions of their empirical analysis is that high school score is a stronger predictor of a student's Final GPA in university than the national mathematics exam score taken at the end of high school. Their results show that neither the high school score nor the national exam score are statistically significant.

In summary, most of the literature on this topic have conducted their experiments using more than one independent variable, applying Multiple Linear Regression and other data analytic techniques. The uniqueness of our study using lies in the use of actual data which are relevant to ATBU current admission policy. Only JME scores and the required 5 credits in SSCE are considered in offering admission to prospective students of the university, no Post-UTME or any other internal exam is needed. Thus, it is pertinent to compare this variable with the university first-year GPA to determine whether there is any significant relationship between them.

III. METHODOLOGY

This section presents the proposed methodology used in fitting the Regression Model for prediction of University Fresher's GPA.

A. *Description of Experiments*

In our proposed prediction model, we used a Simple Linear Regression technique, since we are majorly concerned with only two variables: **d**ependent variable (or variable to be predicted), and **i**ndependent (or predictor or regressor) variable. In this experiment, the variable to be predicted is the GPA and denoted by Y. The unknown variable is the JME score and is denoted by X.

B. *Working Principles of the Proposed Model*

Our main aim in this study is to approximate a simple functional relationship between GPA and JME score by some statistical functions that can allow us to predict the value of one variable from the knowledge of the other variable. The mathematical equation that allows us to predict these values is called a regression equation.

Assume that the regression line of variable Y on variable X has the $\mathbf{Y_i = \alpha + \beta X_i + \varepsilon_i}$ - - - **(1)**
for i = 1, 2, - - -, n, where Yi is the response in the **i**thobservation, Xi is the value of the independent variable, αandβ are constants, while ε_i is the errorof prediction, which is identically andindependently distributed, usually denoted as $\varepsilon_i \sim N(0, \sigma^2)$.

If we take out the error part of equation 1, we have a straight line that we can use to predict values of Y from values of X, which is one of the main uses of regression. Thus, equation 1 would look like this: $Y_i = \alpha + \beta X_i$.

Let a and b be the estimates of α and β. Then the predicted value of Y for a given X when a and b are determined is given by $\mathbf{\hat{Y} = a + bX}$ - - - **(2)**
where a is the y-intercept and b is the slope. X is the independent or predictor variable and Y is the dependent or response variable.

Let's note the difference between the observed value of Y and the estimated value of $\hat{Y}$. While the observed value of Y refers to the Actual GPA of a given Fresher, the estimated value $\hat{Y}$, is the predicted GPA of any Fresher using the regression equation as a means of estimation. Thus, in this research, we use Y to denote the observed GPA of a given Fresher, and $\hat{Y}$ to represent the estimated value of the dependent variable.

Therefore, equation 2 is used as a predictive equation. Substitution of X would provide a prediction of the true mean value of Y for that X. This relationship enables us to compute the value of Y for any given value of X. This does not only enable us to determine the value of Y associated with any given value of X, but it also describes the effect of the change in variable X on the values to be assumed by Y variable.

C. *Assumptions of Linear Regression Model*

1. Both the independent (X) and the dependent (Y) variables are measured at interval or ration level.
2. The relationship between the independent (X) and the dependent (Y) variables is linear.
3. Errors in prediction of the value of Y are distributed in a way that approaches the normal curve.
4. Errors in prediction of the value of Y are all independent of one another.
5. The distribution of the errors in prediction of the value of Y is constant regardless of the value of X.

D. *The Least Squares Method*

The mathematical procedure used to determine the numerical value of the constants 'a' and 'b' in equation 2 is called the Least Squares Method. The constant **a** and **b** are usually derived and obtained as:

$$\mathbf{a = (\sum Y)(\sum X^2) - (\sum X)(\sum XY) / n\sum X^2 - (\sum X)^2}$$

$$\mathbf{b = n\sum XY - (\sum X)(\sum Y) / n\sum X^2 - (\sum X)^2} \text{---} \quad \mathbf{(3)}$$

Substituting **a** and **b** into equation (2) we obtained the Least Squares equation we get equation (4).

$$\mathbf{\hat{Y} = a + bX} \quad \text{- - -} \quad \mathbf{(4)}$$

The general equation (4) is called the regression equation and the constants **a** and **b** are called the regression coefficients. The least squares method finds the values of the constants that minimize thesum of the squared deviations of theobserved values from those predicted by the equation.

E. *Examining the Linear Regression Equation*

There are always four basic assumptions incorporated with Regression Analysis with respect to the error terms. It is assumed that in the regression model $\mathbf{Y_i = \alpha + \beta X_i + \varepsilon_i}$, for i = 1, 2, - - - -, n.

1. ε_i is a random variable with mean zero, and variance σ^2 (unknown). That is, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$. This means that for a given Xi, the differences between Yi and $\hat{Y}$ are sometimes positive, sometimes negative, but on the average are zero. This assumption establishes the estimate-ability of α. That is, only when $E(\varepsilon_i) = 0$ can we get an unbiased estimate of α using the least squares method. $V(\varepsilon_i) = \sigma^2$ means that dispersion or variability of points in the population about the population regression line must be constant.
2. ε_i and ε_j are uncorrelated, $i \neq j$, so that COV $(\varepsilon_i, \varepsilon_j) = 0$. This means that the covariance between the values of the independent variable and the corresponding error terms is zero. This assumption implies that knowing that the error in describing GPA of one (or more) Fresher is positive does not give us any help in determining whether or not the error for another student will also be positive.
3. ε_i is a normally distributed random variable with mean zero and variance σ^2 by (1). Since ε_i is a composite of many factors (such as error of measurements, error in specifying the model, etc.) it is reasonable to expect that many of these factors tend to offset each other so that large values of ε_i are much likely than small ε_i. Simply stated, there is a tendency for errors that occur in many real situations to be normally distributed due to the central limit theorem.
4. The mean of the response, E(*Yi*), at each value of the predictor, *Xi*, is aLinear function of the *Xi*.

F. ***Inferences about the Regression Line***

1) ***Standard Error of the Slope b -Confidence Interval for β:*** The equation 3, for b can be rewritten as:

$$= \sum(X_i - \bar{X})(Y_i - \bar{Y}) / \sum(X_i - \bar{X})^2 \quad --- \quad (5)$$

where

$\bar{X} = (X_1 + X_2 + \cdots + Xn)/n = \sum X_i/n$

$\bar{Y} = (Y_1 + Y_2 + \cdots + Yn)/n = \sum Y_i/n$

In the above expression, one variable is represented by Xi and the other by Yi. The $\bar{X}$ and $\bar{Y}$ are the means of the two set of variables. The terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ are the deviations from the mean as used in computing the variance and standard deviation.

Let us assume the variance of a function F is given as

$F = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$, then

$V(F) = a_1^2V(Y_1) + a_2^2V(Y_2) + \cdots + a_n^2V(Y_n)$, if the Yi are pairwise uncorrelated and the a_i are constants.

If $V(Yi) = \sigma^2$, then $V(f) = (a_1^2 + a_2^2 + \cdots + a_n^2)\sigma^2 = (ai)^2 \sigma^2$. Thus, from the expression b, we have $a_i = \sum(X_i - \bar{X}) / \sum(X_i - \bar{X})^2$,since Xi can be regarded as constants.

Hence, after reduction, the variance of b, is given as

$$\mathbf{V(b) = \sigma^2/\sum(X_i - \bar{X})^2} \quad --- \quad (6)$$

The standard error of b is the square root of the variance. That is, $Se.(b) = \sigma/(\sum(X_i - \bar{X})^2)^{1/2}$ for the unknown.

Therefore, the estimated standard error of b is given by **est.**

$$\mathbf{Se.(b) = S/(\sum(X_i - \bar{X})^2)^{1/2}} \quad --- \quad (7)$$

where S is the estimated σ.

Furthermore, if we assumed that the variance of the observations about the regression line are normal as stated in the assumptions, we can assign 100(1- α)% confidence limits for b by calculating:

$$\mathbf{b \pm t(n-2, 1 - 1/2\, a) / (\sum(X_i - \bar{X})^2)^{1/2}} \quad ---- \quad (8)$$

Where t (n – 2, 1 – ½α) is the percentage amount of distribution with n -2 degrees of freedom.

2) ***Standard Error of the Intercept a: Confidence Interval for α:*** A confidence interval for α (the intercept) and a test of whether or not **α** is equal to some specified value can be constructed in a similar way to that of **β** (the slope). *The estimated standard error of intercept (a) is given as*

$$\mathbf{Se.(a) = (\sum Xi^2/(\sum(X_i - \bar{X})^2)^{1/2} * \sigma} \quad --- \quad (9)$$

3) ***F-test for Significance of Regression:*** We have seen that the Yi are random variables and any function of them is also, a random variable.

We shall consider two particular functions, the mean square due to regression (MSR) and the mean square due to residual variation (MSE). These functions have their own distribution mean variance and moments. It can be shown that their mean values are as follows:

$E(MSR) = \sigma^2 + \sum(X_i - \bar{X})^2.$

$E(MSE) = E(S^2) = \sigma^2$

These expected values suggest how to test H_0: β = 0 versus H_a: β ≠ 0:

- If β = 0, then we would expect the ratio MSR/MSE to equal 1.
- If β ≠ 0, then we would expect the ratio MSR/MSE to be greater than 1.

Since these two variables are independent, a statistical theorem tells us that the ratio

$$\mathbf{F = MSR/MSE = MSR/S^2} \quad --- \quad (10)$$

followsan F-distribution with 1 and (n-2) degrees of freedom provided that β = 0.

In particular case of fitting a straight line, the F-test for regression is exactly the same as the usual t-test for β = 0.This is because F is:

$$MSR/S^2 = b * \sum(X_i - \bar{X})(Y_i - \bar{Y})/S2$$
$$= b * \sum(X_i - \bar{X})/S2$$
$$= (b*(\sum(X_i - \bar{X})^2)^{1/2}/S)^2$$
$$\mathbf{= t^2.}$$

Since the variable F(1, n - 2) is the square of the t(n-2) variable, it gives exactly the same test results.

The advantage of the F-test is that it can be generalized to a test of significance when there is more than one independent variable, while the t-test cannot. However, the t-test has its advantages, in that it can be used to test for value of β other than zero, while the F-test is only appropriate for the null hypothesis Ho: β=0.

F-Test for Simple Linear Regression

To test Ho: β =0 vs. Ha: β ≠0, we use the following procedure:

1. Compute the F-test statistic: *F* = MSR/ MSE, that follows an $F_{1,\, n-2}$ distribution under *Ho.*
2. For a two-sided test with significance level α,
   If $F > F_{1,\, n-2,\, 1-\alpha}$ then reject Ho;
   If $F \le F_{1,\, n-2,\, 1-\alpha}$ then accept Ho.
3. The exact p-value is given by $Pr(F_{1,\, n-2} > F)$.

T-Test for Simple Linear Regression

1. Compute the test statistic *t* = b/Se.(b)
2. For a two-sided test with significance level α,
   If $t > t_{n-2,1-\alpha/2}$ or $t < t_{n-2,\alpha/2} = -t_{n-2,1-\alpha/2}$ then reject Ho.
   If $-t_{n-2,1-\alpha/2} \le t \le t_{n-2,1-\alpha/2}$ then accept Ho.
3. The p-value is given by p=2 x (area to the left of t under a t_{n-2} distribution) if t < 0, p = 2 x (area to the right of t under a t_{n-2} distribution) if t α ≥0.

4) ***Standard Error of $\hat{Y}$:*** We can also show that the the regression equation $\hat{Y} = \bar{Y} + b(X - \bar{X})$ where both $\hat{Y}$ and b are subject to error, which will influence $\hat{Y}$. If $\bar{Y}$ and b are uncorrelated random variables, then the variance of the predicted mean value of Y, $\hat{Y}_k$ at a specific value of X_k of X is given by: $\mathbf{V(\hat{Y}_k) = V(\bar{Y}) + (X_k - \bar{X})V(b) =}$

$$\sigma^2/n+(X_k-\bar{X})^2\sigma^2/\sum(X_i-\bar{X})^2 \quad - - - \quad (11)$$

Hence, **est. Se.** $(\hat{Y}_k)= S*(1/n + (X_k - \bar{X})^2/\sum(X_i - \bar{X})^2)^{1/2}$ **(12)**

The variance and standard error shown above apply to the predicted mean value of Y for a given $\mathbf{X_k}$. Since the actual observed value of Y varies about the true mean value with variance σ^2 (independent of the $V(\hat{Y})$, a predicted value of an individual observation will still be given by Y, but will have variance $\sigma^2 + V(\hat{Y}_k) = \sigma^2(1 + 1/n + (X_k - \bar{X})^2/\sum(X_i - \bar{X})^2)$ **(13)** with the corresponding estimated value obtained by inserting S^2 for σ^2.

G. *Standard Error of Estimate*

The standard error of estimate is one of the most useful measures of goodness of fit in regression analysis. It is denoted by the symbol **Se**. Mathematically; these are the predicted values of Ysubtracted from the actualobserved values of Y.

Standard error of estimate is usually computed using the equation:

Se = square root of $(Y_i^2 - aY_i - bX_iY_i) / n - 2 =$ **sqrt (SSE / (n-2)) - - -** **(14)**

H. *Correlation Coefficient (R) and Coefficient of Determination (R^2)*

A correlation coefficient (R) is a measure of the relationship between two variables. It describes the tendency of two variables to vary together (covary). That is, it describes the tendency of high or low values of one variable to be regularly associated with either high or low values of the other variables.

The mathematical formula for computing R is:

$$R = n\sum XY - (\sum X)(\sum Y / Sqrt([n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]) \quad - - - \quad (15)$$

where:

n = The number of paired scores observed
X = Independent variable (JME Scores)
Y = Dependent variable (GPA Scores)
$\sum Xi$ = Sum of the raw JME scores
$\sum Yi$ = Sum of the raw GPA scores
$\sum XiYi$ = Sum of the products of each X line by each Y line
$\sum Xi^2$ = Sum of the squares of each JME score
$\sum Yi^2$ = Sum of the squares of each GPA score
$(\sum Xi)^2$ = The squares of the total sum of JME scores
$(\sum Yi)^2$ = The squares of the total sum of GPA scores.

The value of R is such that $-1 \leq R \leq +1$. That is, the linear coefficient (R) may be positive or negative. If R is positive, Y tends to increase with X (meaning the slope of the least squares line is positive), while if R is negative, Y tends to decease with X (i.e., the slope is negative).

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak **[17]**.

The Coefficient of determination, R^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. R^2 is such that $0 \leq R^2 \leq 1$, and denotes the strength of the linear association between X and Y.

Hypotheses Test on the Correlation Coefficient

To test the hypothesis Ho: $\beta = 0$ vs. Ha: $\beta \neq 0$, we use the following procedure:

1. Compute the sample correlation coefficient R
2. Compute the test statistic $t = R(n-2)^{1/2}/(1-R^2)^{1/2} = R*Sqrt(n-2)\ /sqrt(1-R^2)$, Which under Ho, follows a t distribution with n-2 df

For a two-sided level α test

if $t > t_{n-2,1-\alpha/2}$ or $t < -t_{n-2,1-\alpha/2}$ then reject *Ho*.
If $-t_{n-2,1-\alpha/2} \leq t \leq t_{n-2,1-\alpha/2}$, then accept *Ho*.

3. The *p*-value is given by
$p = 2 \times$ (area to the left of t under a t_{n-2}distribution) if $t < 0$
$p = 2 \times$ (area to the right of t under a t_{n-2}distribution) if t α ≥ 0.

Here we assume an underlying normal distribution for each of the random variables used to compute *R*.

I. *ANOVA Test for Regression*

An alternative method used for testing whether a linear relationship exists between X and Y is the used of Analysis of Variance (ANOVA). This is computed with following equation:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad --- \quad (16)$$

Equation 16 may be rewritten as SST = SSR + SSE, where SS is the notation for sum of squares and T, R and E are notations for Total, Regression model, and Error, respectively. With these formulas, information about the Mean Square Error (MSE) and the Mean Square Regression (MSR) can be computed respectively as MSR = SSR/1 and MSE = SSE/ (n - 2) where 1 and (n – 2) are the degrees of freedom for the two variables.

The Significance of the regression can also be tested by comparing the ratio of the regression and residual mean squares with the F-distribution having 1 and (n -2) degrees of freedom.

For a simple regression, the format for calculations of ANOVA are displayed as in table II:

TABLE II
ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

Source of Variation	**Sum of Squares**	**Degree of Freedom**	**Mean Squares**	**F-value**
Regression	SSR	1	SSR/1	MSR/MSE
Error (Residual)	SSE	n - 2	SSE/(n – 2)	
TOTAL	SST	n - 1	SST/(n -1)	

J. *Examination of Residuals*

The residuals (errors of prediction) are defined as the n differences $\mathbf{\varepsilon_i = Y_i - \hat{Y}_i}$, - - - **(17)** for i = 1, 2, . . ., n, where Yi is an observation and $\mathbf{\hat{Y}_i}$ is the corresponding fitted value obtained by used of the fitted regression equation. This is the difference between what is actually observed and what is predicted by the regression equation – given the amount which the regression equation has not been able to explain.

Examination of residuals provides information relevant to two basic types of observations. Firstly, it may indicate lack of linearity and provides guidance towards most appropriate modifications.For example, the pattern of residualsmay indicate the need for adding terms to the equation.

The second type of observation which may be answered by examination of residuals is whether the assumptions about the errors are met (see section E). An examination of residuals is therefore, an important part of regression analysis because it helps to detect any inconsistency between the data and the postulated model.

IV. RESULTS AND DISCUSSION

In this section, we analyzed and discussed the implemented results of the proposed predictive model.

A. *Simulation Environment*

The work was implemented using MATLAB version R2017a (9.2.0.538062) installed on an intel Pentium N3540 Processor, 2.16GHz CPU with a 500GB hard disk and 4GB RAM. For the purpose of comparison, simulations were also carried out in Minitab 19.2020 (64-bit) and Microsoft Excel data analysis.

B. *Data Collection*

Data collection was one of the first steps taken in the course of carrying out this research. The GPA and JME scores data from 2015 to 2019 were collected from Abubakar Tafawa Balewa University academic office, Bauchi. A statistical sampling technique called systematic sampling method was used in collecting the data. This technique was chosen among many techniques such as simple random sampling, stratified random sampling, etc., for its ease and accuracy of applications.

According to **[18]**, a systematic sample is a simple random sample of one cluster from a population of K cluster units. That is, to select asample of n units, we take a unit at random from the first k units, and every k^{th} unit thereafter. Generally, K = N/n where N is the total number of members in the population (population size) and n is the desired population.

For this research, N = 931, n = 100 and K = 9. The first unit drawn was a student numbered 2 in the list given to me and the subsequent units are 11, 20, 29, etc. Thus, the selection of the first unit determines the whole sample to be taken.

C. *Presentation and Analysis of Results*

The results for the experiment and simulation of the proposed Regression model are described in this section.

From the scatterplot in Figures 1a. and 1b, we see that the smoothing line is not pretty linear. This may be due to outliers and many other factors that can affect student's academic performance as stated in (**[19], [22]**). In this study, we are not interested in removing outliers, because they are part of the population that we are studying. The outliers cover for the Bauchi State indigene students that are admitted based on their excellent SSCE performance and therefore promoted directly into 100-level.

The overall pattern of the plotted points clearly shows that the relationship between GPA and JMEscores is very weak but positive with a correlation R = 0.1372.

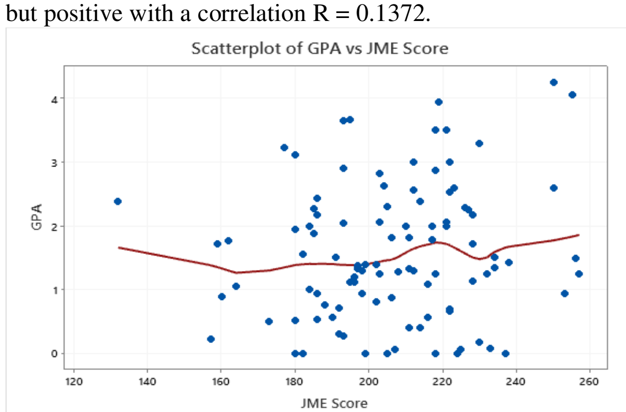

Fig. 1a: Scatter diagram of GPA and JME Scores with Minitab

Fig. 2b: Scatter diagram of GPA and JME Scores with MATLAB

1) ***The Regression Coefficients:*** The method of least squares described in section D is used to estimate the coefficients of regression and to derive the least squares prediction equation for the JME Score and GPA data from 2015 to 2019. To obtain the estimates for the data, we calculate:

The sum of X (JME Scores) = 226 + 218 + 204 + - - - + 193 = 20612

The sum of Y (the GPAs) = 2.29 + 2.88 + 2.63 + - - - - + 2.90 = 158.09

Using these on the formulas, the following values for the regression coefficients were obtained: 0.296 being an estimate for α, and 0.00623 as estimate for β. Thus, a =0.296, b=0.00623.

The estimate, a = 0.296 tells us that the regression line will intersect the Y axis at the value of 0.296, and the b =0.00623 coefficient implies that when X (the JME score) increases by one point, then Y (the GPA) increases by 0.00623. The positivity of 'b' also indicates that the relation of between GPA and JME scores is in the positive axes of the regression line.

2) ***Prediction using the Regression Equation:*** Using the values of **a** and **b** obtained in section D on the least squares regression equation relating GPA and JME, we have

$\hat{Y}$ = a + bXi = 0.296 + 0.00623(Xi) - -- **(18)**

From Figure 2a, we could see outliers at the top right corner of the fitted line plot. Also, the p-value for our regression is 0.173. This is greater than the significance level α = 0.05. Generally, P-value > α, means there is not enough evidence to conclude that the regression model explains variation in the response. This may be strongly linked to the effect of the outliers and other factors.

In the regression output for Minitab, we can see **S** in the summary of Model section, right next to R-squared as displayed in Figure 2a. S is known both as the standard error of the regression and as the standard error of the estimate. S is also an overall measure of how well the model fits the data.

In our case, S is 1.0571, which tells us that the average distance of the data points from the fitted line is about 1.06% GPA. And the R-sq means that 1.9% of the variation of our regression model is due to JME score, while R-sq(adj) tells us that only 0.9% of the variation explained by the JME score actually affects the GPA score. Note that both R-sq and the R-sq(adj) give us an idea of how many data points fall within the line of the regression equation.

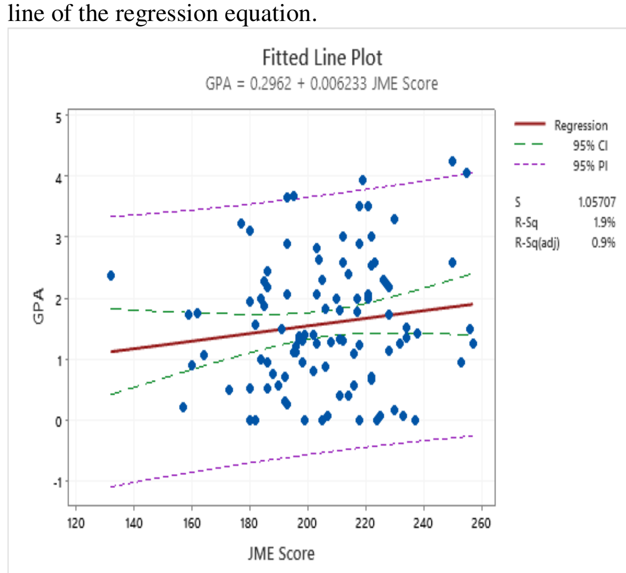

Fig. 2a. Fitted Line plots with Minitab

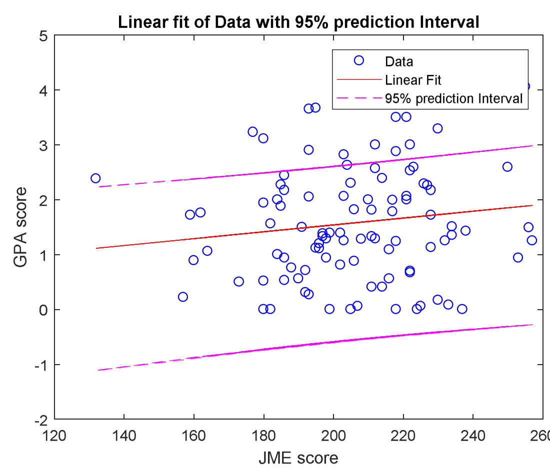

Fig. 2b. Fitted Line plots with MATLAB

The regression line enables us to make prediction of any value of GPA, given a particular JME score. For example, if we wish to predict GPA performance of a student with JME score of 250, the regression equation would be $\hat{Y} = 0.296 + 0.00623(250) = 1.85$. That is, given a JME score of 250, the predicted GPA is 1.85. Hence our best estimate for the mean GPA of all university Freshers having JME score of 250 is also $\hat{Y} = 1.85$.

This regression equation for example, will predicts that the GPA for students with JME score of 90 will be $\hat{Y} = 0.296 + 0.00623(90) = 0.86$. But since students with this low score are usually not admitted in the university, there is no way of knowing what GPA they might achieve.

Similarly, the equation we derived can predict the GPA for a student with a JME score of 900 to be $\hat{Y} = 0.296 + 0.00623(900) = 5.90$ which is outside the sample data drawn, and not even possible since the maximum score a student can achieve in JAMB is 400 (100 in each of the 4 subjects, with English and Math compulsory).

From equation (18), the sign of the coefficient b is positive (that is, +0.00623), indicating that the relationship between JME score and GPA is positive.

Clearly, special care must be taken in this study when attempting to predict outside the range of JME scores which fall outside the range of past experience (historical data used), since it may be that these values cannot be represented by the same equation. In this study, prediction of GPA should be taken from the JME score range of 100 to 400.

3) ***Standard Error of Prediction:*** The standard error of estimate, denoted Se was describe in section G and is obtained as Se = 1.0571. This indicates that the standard deviation of the GPA predicted from the JME scores is 1.0571. In other words, the standard deviation of the Y-scores when X, say 226 is 1.0571, and it is also 1.0571 for X=218, and so on.

We used Residual Plots to examine the goodness of model fit in regression and ANOVA. Examining the residual plots will help us to determine if the least squares method assumptions are met (see section E). Basic Minitab residual plots depicted in Figure 3 established that the residuals are normally distributed.

1. **Normal Probability Plot of residuals**. The normal plot of residuals has verified the assumption that the residuals are normally distributed (except for the few outliers pointed out earlier).
2. **Residuals Versus Fitted Values**. This plot shows a non-systematic pattern of residuals on both sides of 0 as can be seen below. The residuals versus fits plot is used to verify the assumption that the residuals have a constant variance.
3. **Histogram of residuals**: This showed that the data are skewed to the right.
4. **Residuals versus order of data:** This plot is used to verify the assumption that the residuals are uncorrelated with each other.

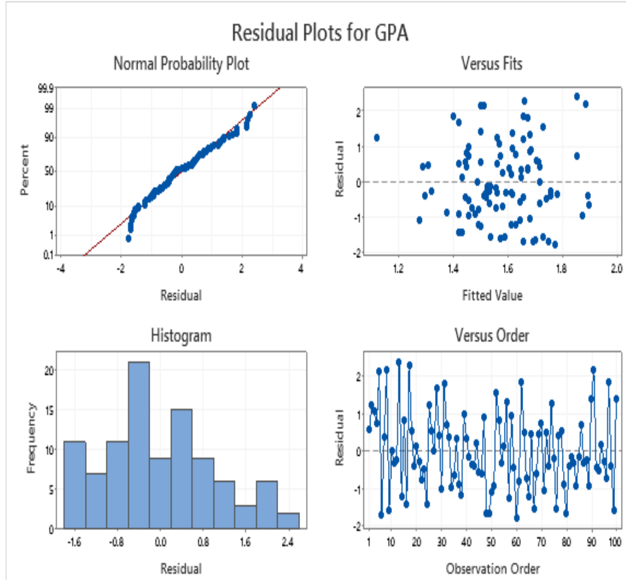

Fig. 3. Residual Plots

In this study of standard error of estimate, we note that a prediction is made of the mean score on a second variable (Y) for a whole multitude of individuals who have a given score on the first variable (X). For example, for the multitude of Freshers scoring 226 on the X variable, there is a predicted Y-score of (0.296 + 0.00623*226) =1.70398, with a deviation of 1.0571, and the assumption that a GPA of 1.85 would be predicted for a JME score of 250 does not mean that every individual who score 250 on the exam will receive a GPA of 1.85. But rather, it means that 1.85 is the predicted mean GPA for a sub-population of Freshers who score 250 on the exam.

Generally, we can say that there is a variability about the predicted mean GPA. Some students with a score of 250 will get higher GPAs and some will get lower GPAs. Therefore, in prediction like this, it is very important to get an estimate of the amount of variability observed in the GPA score for persons with a given JME score. Thus, 1.85 determines the variability to be expected in the GPAs of all the students who had a JME score of 250.

4) ***Inferences about the Regression Line:*** As we described in section F, the statistical inferences about the regression line were computed and the following results obtained:

Using equation 7, estimated Se.(b) = 0.00455,
Using equation 9, estimated Se.(a) = 0.94285
Using equation 10, calculated F = 1.88018
Using equation 12, estimated Se($\hat{Y}_k$) is obtained as

$Se(\hat{Y}_k)=1.0571*[1/100+(X_k -206.12)^2/54082.56]^{1/2}$

This estimated $Se(\hat{Y}_k)$ can be used to test for different values of Xk. For example, for $X_k = 226$, the est. $Se(\hat{Y}_k) = 0.13901$. Correspondingly, when $X_k =218$, the est. $Se(\hat{Y}_k) = 0.11870$, etc.

The standard deviation of b (Sb) = 0.00455 was used to test the estimate of the slope of the regression line (b or β).

To infer from the sampling distribution for our statistic b, whether the sample data provide sufficiently strong evidence that higher JME scores are related to GPA in the population, we set up our null and alternative hypotheses as follows:

1. **Null hypothesis: Ho:** $\beta = 0$ (there is no relationship between GPA and JME scores)
2. **Alternative hypothesis: Ha:** $\beta \neq 0$ (there is a relationship between GPA and JME scores).

Assume the commonly used significance level α=0.05, for a two-sided test, and we compute $1 - \alpha/2$ or $1 - 0.05/2=0.975$. Then our t(98, 0.975) = 1.984 as obtained from the t-distribution table. Also, the 95% confidence limits for this b is $b \pm t(98, 0.975) * S/\text{sqrt}(\sum (X_i - \bar{X})^2)$. That is, since the coefficient b follows a t-distribution with (n-2) degrees of freedom (d.f.) and a standard deviation S_b, the desired interval is $b - t(n\text{-}2, 1 - \alpha/2)*S_b \leq \boldsymbol{\beta} \leq b + t(n\text{-}2, 1- \alpha/2)*S_b$.

With our n = 100, $t(n\text{-}2, 1 -1/2\alpha) = t(98, 0.975) = 1.984$ and Sb =0.00455, hence the 95% confidence interval would be $0.00623 - (1.984)*(0.00455) \leq \beta \leq 0.00623+(1.984)*(0.00455)$ $= -0.00279 \leq \beta \leq + 0.01526$.

In other words, the true value of β lies in the interval (-0.00279 to 0.01526) and this statement is made with a 95% confidence.

Writing the above as Ho: $\beta = 0$, H_a: $\beta \neq 0$ and evaluating t = b /Se.(b), we obtained t = 0.00623 /0.00455 = 1.371. And since $|t|$ = 1.371 is less than the appropriate critical value of t(98, 0.975) = 1.984, then Ho: β=0 cannot be rejected. Hence, there is no relationship between GPA and JME scores. Thus, we conclude that the regression line does not strongly seem to improve our ability to estimate the dependent variable (GPA) from the JME scores.

5) Correlation Coefficient and Coefficient of Determination: The correlation coefficient (R) as described in section H., indicates the direction and strength of the relationship between GPA and JME.

Our correlation coefficient R = 0.1372 means that the data points are scattered near the line with a positive slope. This shows a poor linear relationship between the data since coefficient value is numerically close to zero.

Valid interpretation of the coefficient can be made with the coefficient of determination, which is just the square of the R value. R^2 is the summary measure of goodness of fit frequently referred to in the literature**.**

Our R^2 = 0.0188, represents the proportion of GPA accounted for by its linear relation with JME scores. That is, the R of 0.1372 tells us that 0.0188 or 1.9% of the GPA variance can be accounted for by its correlation with the JME scores. Therefore, to predict GPA from JME scores we recognize that we would be able to account for only 1.9% of GPA. The other 98.1% is independent of JME score that cannot be accounted for. This may be due to many factors that may affect students' performance such as the student health, reading habit and reading plan, self-discipline, lecturers' teaching and training method, environmental factors, or family financial status as stated in (**[2], [6]**, **[12], [15], [19], [22],** etc.).

Conclusively, having the correlation of R=0.1372, it would be possible to make predictions of scholastic achievements of university Freshers' first-year GPAs from the JME scores of similar groups if factors that can affect a student's performance are taken care off.

6) ***ANOVA Test for Regression:*** Based on the formulas in table II., the Analysis of Variance for the regression of GPA on JME score is as shown in table III.

TABLE III
ANOVA TABLE FOR REGRESSION OF GPA AND JME SCORES

Source of Variation	Degree of freedom	Sum of Squares	Mean Squares	F-value	F-critical
Regression	1	2.101	2.101	1.880	3.938
Error (residual)	98	109.505	1.117		
TOTAL	**99**	**111.606**			

Usually, the ANOVA table is used to hypothesize the null hypothesis that the slope of the regression line relating GPA and JME scores does not equal to zero (Ha: $\beta \neq 0$). In other words; the hypothesis that there is no dependence of Y on X.

This is done by comparing the ratio of the regression and residual mean squares with the F-distribution having degrees of freedom of 1 and n-2 respectively. In our case, the ANOVA gives the calculated F as F-calculated $=MSR/S^2 = MSR/MSE = 2.101/1.117 = 1.880$.

For a simple linear regression, the statistic MSR/MSE has an F-distribution with degrees of freedom as follows: F(DFR,DFE) = F(1, n-2) = F(1,98) for n=100, where DFR = degree of freedom for the sum of squares of Regression (SSR) and DFE = degree of freedom for the sum of squares of Error (residuals) or (SSE).

This result of F-calculated agrees perfectly with the proposition that $F = t^2$ as stated in section F(3). This is because $t^2 = (1.371)^2 = 1.880$ (value of F).

The ANOVA table from Minitab shows the results for F-value and P-value as can be seen in table IV.

TABLE IV
ANOVA TABLE FROM MINITAB

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.101	2.10091	1.88	0.173
Error	98	109.505	1.11740		
Total	99	111.606			

The critical value of F for this F-value is obtained from the F-distribution table as 3.938. This means that F-value is less than F-critical (1.88 < 3.94). But, according to **[20]**, since the observed F-value is less than the critical F-value, the contribution of variance of 1.9% to the achievement of students by JME scores is not significant. Thus, the null hypothesis cannot be rejected.

We therefore, conclude that the linear relationship does not help in explaining the variation in GPA. This is exactly in agreement with the test of the slope in section IV. (4). Also, our P-value = 0.173 is greater than the significance level ($\alpha = 0.05$), implying that the null hypothesis cannot be rejected.

7) ***Examination of the Residuals:*** To obtain the error term $\mathcal{E}_i$ for our data, the formulas $\mathcal{E}_i = Y_i - \hat{Y}_i$ and $\mathcal{E}_i^{\ 2} = \sum (Y_i - \hat{Y}_i)^2$ were used respectively. This gives the error term $\mathcal{E}_i$ = 0.585, 1.225, - - - -, 1.401 and the square error $\mathcal{E}_i^2$ = 0.343, 1.501, - - - - - -, 1.962. In this case,
the sum of $\mathcal{E}_i$ = 0.585 + 1.225 + - - - - + 1.401 = 0.000, while the sum of $\mathcal{E}_i^{\ 2}$ = 0.343 + 1.501 + - - - + 1.962 = 109.505.

A method to evaluate the regression model is by analyzing the error terms (residuals) by plotting them on a graph. Figure 3 shows that the residuals appear to behave randomly and show no particular pattern, which means that the model is good, though not significant.

D. ***General Observation on Results***

A regression model is fitted for the prediction of first-year GPA from JME scores of university Freshers. The following findings were made with respect to the objectives of the research:

1. Simple Linear Regression is not good for fitting an accurate model for comparing JME scores and first-year GPA of University Freshers, owing to several factors that can affect students while in the University.
2. The residual plots of our model showed that all the assumptions stated for error of estimate are met, namely:
   a. The residuals are normally distributed.
   b. The residuals are random, with no specific pattern on both side of the zero.
   c. Histogram of residuals show that the data are skewed to the right.
   d. Residuals vs order of data plot indicates that the residuals are uncorrelated with each other.
3. The measure of degree of association between the two variables is only moderate even though positive.

The findings of this study; that there is no significant relationship between JME scores and University students' GPA are in agreement with the following researchers' findings which reported among others, that the correlation between JME scores and the students' performance at first year proved to be non-significant (**[10]; [13];[14]; [15];** etc.).

Also, recently, the author of **[21]** corroborated this in their work in which the results showed that candidates' scores in JAMB do not actually represent their true scores or ability.

V. CONCLUSION AND RECOMMENDATION

This project was confined to the prediction of students of Abubakar Tafawa Balewa University Bauchi, using statistical data collected from the academic office.

The review of current methods of selection on the basis of prediction showed that there is an absence of corrective means for predicting those applicants who are deemed to be more successful in JME for admission into the University.

Data collected were analyzed by the Least Squares Method and residual analysis. Chapter four deals with discussion of the results obtained, in which some comparisons and observations on hypotheses were made. The results of hypotheses, ANOVA, and prediction led to concluding that there is no strong relationship between GPA and JME scores.

In other words, the linear relationship between the two variables does not help in explaining the variation in GPA, in that sometimes students with high intelligence tend to make low grades when in the University, majorly owing to various factors.

Therefore, based on the findings of this study, we conclude that JME scores alone are not predictors of first-year students' GPA in the university. The implication of this is that much emphasis should not be placed on the JME scores alone. Rather, other indicators like university internal exams, and the weighted aggregate score in the 5-SSCE subjects required as pre-requisite for particular course should be taken into consideration before admitting students.

A. ***Recommendations***

Having made some conclusions based on the results of the analysis and prediction, it is necessary to make some recommendations to the University for the effective implementation of this predictive model in the future. These include:

1. Corrective measures such as those studied in this research should be used in holistic analysis of data for prediction of students' academic performance.
2. The university should provide adequate facilities for students in term of accommodation on campus and serene environment, as this will enhance theirconcentration, and safeguard a lot of students' retrogression in performance when admitted into the University.
3. Students should be encouraged through counselling and career guidance to maintain their academic performance. In this realm, Federal Government could be called upon to support the University system, both financially and materially.
4. The selection process used by the University should be enhanced to guarantee unbiased prediction during admission process.
5. The findings of this study call for an in-depth study on the reliability and validity of the JME scores (entrance examination results) that is being used as a major criterion to University admission in ATBU and other higher institutions in the country at large.

Finally, we would not be wrong at all, if we predict that students with above-average JME scores may achieve low score in GPA, and those with below-average JME scores can get high score in GPA, and rightly conclude that GPA and JME scores are only moderately related since there is tendency that students who score high in one can score low in the other and vice versa.

In the future, we will employ some of the factors that affect students' academic performance and apply multiple regression to experiment with.

ACKNOWLEDGMENT

I wish to acknowledge my friend and beloved brother, Dr. Titus Ketkukah for inspiring me to carry out this research. Thanks a lot for your immeasurable advice and encouragement throughout the study. May the Almighty God bless you richly. Amen.

I would like to thank Dr. Nentawe Gurumdimma for supervising and providing useful comments that have enriched the research work. Your brilliant guidance, despite your tight schedules will always be treasured in my heart.

I would like to leave warm words of gratitude to my M.Sc. thesis supervisor, Professor Souley Boukari, for his insightful guidance, invaluable advice, and constructive criticism during the project. His academic and research expertise has helped so much in improving my research skills.

Finally, but not at all the least, I wish to sincerely acknowledge my family (wife and children) for their moral and financial support throughout this study.

REFERENCES

[1] D. Ary and L. C. Jacobs, *Introduction to Statistics: Purpose and Procedures*, 76th ed., Holt, Rinehart & Winston, USA, 1976.

[2] D. Baneres, M. E. Rodriguez-Gonzalez, and M. Serra, "An early feedback prediction system for learners at-risk within a first-year higher education course," *IEEE*, vol. 12, issue 2, pp. 249–263, 23 Apr. 2019.

[3] A.S.R Abdallah, and M.O.M Mohammeda, "Application of regression methods to investigate the factors influence on students' grade point average. Contents lists available at GrowingScience, Management Science Letters 11 (2021), 11, pp. 1-8, Sep. 2020. www. Growingscience.com/msl.

[4] F. Widyahastuti, and V. U. Tjhin, "Predicting students' performance in final examination using linear regression and multilayer perceptron," in International Conference on Human System Interaction *(HIS),* 2017, p. 188-192.

[5] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: a systematic literature review. ITiCSE 2018 Companion: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, July 2018 p. 175–199. https://doi.org/10.1145/3293881.3295783

[6] M. M. Sulphey, N. S. Al-Kahtani, and A. M. Syed, "Relationship between admission grades and academic achievement,". The International Journal of Entrepreneurship and Sustainability Issues, vol. 5, issues 3, pp. 648-658, 1 Aug. 2018. http://doi.org/10.9770/jesi.2018.5.3(17).

[7] E. Alyahyan, and D. Düştegör, "Predicting academic success in higher education: literature review and best practices,". *International Journal of Educational Technology in Higher Education,* vol. 17, issues 3, pp. 1-21, 26 Mar. 2020. https://doi.org/10.1186/s41239-020-0177-7.

[8] B. B. Aina, "UTME and post-UTME scores as predictor of academic performance of university undergraduates of Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria". International Journal of Education and Evaluation ISSN 2489-0073, vol. 3 issue 1, pp. 37-43. 2017

[9] B. S. Aribisala, O. A. Dosunmu, M. A. Akanbi, and I. O. Shittu, "Comparative analysis of university matriculation examination and post university matriculation examination admission models in Lagos State University". Ethiopian Journal of. Science & Technology, vol. 11 issue 2, pp. 129-143.2018. https://doi.org/10.4314/ejst.v11i2.4

[10] S. M. Hassan, and M. S. Al-Razgan, "Pre-university exams effect on students' GPA: a case study", in IT department. Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia., p. 127-131.Available online at www.sciencedirect.com.

[11] R. Dorta-Guerra, I. Marrero, B. Abdul-Jalbar, R. Trujillo-Gonzalez, and N. V. Torres, "A new academic performance indicator for the first Term of first-year science degree students at La Laguna University -a predictive model", FEBS Openbio 9: Education Article. Published by FEBS Press and John Wiley & sons ltd, vol. 9, issue 9, pp. 1493-1502. 30 Jul. 2019. https://doi.org/10.1002/2211-5463.12707.

[12] F. Egberha, and J. H. Osarumwense, "University Matriculation Examinations as Predictor of First Year Academic Performance of Students in Public Universities in Edo State". Journal of Education and Practice, vol. 10, issues 15, pp. 95 -101. 31 May 2019. DOI: 10.7176/JEP/10-15-12.

[13] B. N. Nweze, "University Matriculation Entrance Examination Scores in Chemistry as Predictors of Achievement in Chemistry-Based Courses in Public Universities in Enugu State Enugu State University

of Science and Technology, Enugu, State, Nigeria", in 2019 ASEE, annual conference & exposition, paper #25319, p. 1-19.

[14] J. A. Odukoya, S. I. Popoola, A. A. Atayero, D. O. Omole, J. A. Badejo. T. M. John, and O. O. Olowo, "Learning analytics: Dataset for empirical evaluation of entry requirements into engineering undergraduate programs in a Nigerian university", Published by Elsevier Inc, vol. 17, pp. 998-1014. Apr. 2018. https://doi.org/10.1016/j.dib.2018.02.025.

[15] P. L. Silva, L. C. Nunes, C. Seabra, A. B. Reis, and M. Alves, "Student selection and performance in higher education: admission exams vs. high school scores", Journal of *Education Economics*, vol. 28, issue 5, pp. 437-454. 26 Jun. 2020. https://doi.org/10.1080/09645292.2020.1782846.

[16] S. J. H. Yang, O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, H. Ogata, and A. J. Q. Lin, "Predicting students' academic performance using multiple linear regression and principal component analysis". Journal of Information Processing, vol. 26, pp. 170-176. Feb. 2018. https://doi.org/10.2197/ipsjjip.26.170.

[17] K. Storm, Introduction to construction statistics using Excel. Industrial Process Plant Construction Estimating and Man-Hour Analysis, eBook in ScienceDirect, p. 1-21. Available online 31 May 2019. https://doi.org/10.1016/B978-0-12-818648-0.00001-6.

[18] A. Donald, C. J. Lucy, and K. S. Christine, *Introduction to Research in Education, 8th Edition*, (pdfdrive.com), Wadsworth, Cengage Learning. 12 Feb. 2010.

[19] H. T. T. Le, H. T. T. Nguyen, T. P. La, T. T. T. Le, N. T. Nguyen, T. P. T. Nguyen, and T. Tran, "Factors Affecting Academic Performance of First-Year University Students: A Case of a Vietnamese University". International Journal of Education and Practice, vol. 8, issue 2, pp. 221-232, 27 Feb. 2020. https://doi.org/10.18488/journal.61.2020.82.221.232.

[20] B. Rosner, *Fundamentals of Biostatistics, 8th edition,* (pdfdrive.com), Boston: Brooks/Cole, Cengage Learning. Available in www.cengage.com/highered, ISBN# 978-1-305-26892-0.

[21] H. O. Etaga, E. Njideka, and K. F. Aforka, "The Relationship between Post UTME and UTME Scores of Students Admitted into Universities in Nigeria". African Journal of Social Sciences and Humanities Research, *vol. 3, issue* 5, pp. 88-96. 24 Sep. 2020.

[22] S. O. Olatunji, D. O. Aghimien, A. E. Oke, and E. Olushola, "Factors Affecting Performance of Undergraduate Students in Construction Related Disciplines". Journal of Education and Practice, vol. 7, issue 13, pp. 55-62. 2016. Available in https://iiste.org/journals/index.php/JEP/article/view/80621