Predicting Students' Final Grade in Mathematics Module using Multiple Linear Regression

Hazlina Darman, Sarah Musa, Rajasegeran Ramasamy, Raja Rajeswari

Abstract: ***Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. In this paper, multiple linear regression model is developed to predict the students' score in Final Exam using their assessments' score. The response variable in this model is the students' score in Final Exam and the predictor variables are the assessment components (Test 1 and Test 2). The data were collected from a group of students in School of Actuarial Science, Mathematics, and Qualitative Study (SOMAQS), Asia Pacific University of Technology and Innovation (APU), Malaysia. In this research, a regression model has been developed with the aid of Statistical Package for Social Sciences (SPSS) analysis tool. The graphical representations and tables are presented to illustrate the models.***

Keywords: Multiple Linear Regression, Students' Performance, Learning Analytics, SPSS, Response, Variable, Predictor Variable, Correlation

I. INTRODUCTION

All students who enrolled in Diploma in Accounting, Diploma in Finance, Diploma in Business IT, and Diploma in Business Administration in APU are required to take Business Statistics (BSTAT) module. This module consists of basic statistics topics such as correlation, regression, probability distribution, sampling distribution and hypothesis testing, and also time series. This module consists of 60% coursework and 40% exam where the passing marks is 50%. The failure rate for this module is considered quite high. A few meetings have been conducted between the module lecturers and academic leader to figure out the solutions of this problem. Thus, the purpose of this paper is to develop a regression model to predict the students' final exam so that the early actions can be taken towards the potential failed students.

Prediction of student academic performance by using different approaches has become a main topic in many different areas. These approaches will help the lecturers or teachers to predict the number of students who will pass or fail in certain modules, prior to the official results. In other words, lecturers or teachers can identify the specific students who have the tendency to fail their modules. Therefore, some early preventionscan be implemented to help those students to get better results in that particular modules.

Revised Manuscript Received on January 19, 2019.

Hazlina Darman, School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

Sarah Musa, School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

Rajasegeran Ramasamy, School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia .

Raja Rajeswari, School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

The preventions can include: forming a few groups of potential failed students and conducting consultations or extra classes for them, giving more fundamental exercises for them to do as a revisions, and etc.

In the previous researches, multiple linear regression has been the most commonly and widely employed for this purpose. Karamazova et.al (2017) has developed simple and multiple linear regression model to analyze and compare the final grades of students from two universities. Same technique were used by Khan and Al-Zubaidy (2017) to predict the students' performance in academic and military learning environment. Green (2005) presented in his paper a set of linear regression model for three mechanical engineering courses to predict students' final exam scores based on the students' scores in mid-term test and quizzes. In research done by Yousuf (2000), they developed a multivariate linear regression model to predict students' academic performances in Computer Science and Engineering Technology programs. A new set of multivariate linear regression models are developed by Huang and Fang (2010) to predict the student academics' performance in Engineering Dynamics Course. Adeyemi (2008) used *z*-test, correlation analysis and multiple regression to predict students' performance in Senior Secondary Certificate Examinations based on their performance in Junior Secondary Certificate Examinations in Ondo State, Nigeria. Some other techniques were also used such as linear and logistic models (Ayan and Garcia, 2008), neural networks (Imbrie et.al, 2008), Bayesian networks (Nghe et. al, 2007), decision trees (Thomas and Galambos, 2004), generic algorithm (Minaei-Bidgoli, 2003), machine learning algorithm (Kotsiantis et.al, 2003)) and data mining techniques under classification (Adhatro et.al, 2013). Three predictive models have been developed by (Ibrahim and Rusli, 2007) using SAS Enterprise Miner, which are artificial neural network, decision tree and linear regression.

The objective of this study is to develop a multiple linear regression model to predict students' exam scores based on their performance in coursework, Test 1 and Test 2. The predictor variables of this model are Test 1 and Test 2 scores, and the response variable is the exam score. The data are collected from the previous intake of students, and SPSS software are used to get the predicted exam scores. The accuracy and the validity of the model are analyzed and the comparison between actual exam marks with the predicted exam marks are done.

II. METHOD AND MATERIALS

The hypotheses of this study are:

H_o: There is no significant relation between assessment scores (Test 1 and Test 2)and Final Exam scores.
H_1: There is asignificant relation between assessment scores (Test 1 and Test 2) and Final Exam scores.

A multiple linear regression analysis is carried out to predict the values of a response variable *y*, students' exam marks. The model is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where
X_1: Test 1'sscore
X_2: Test 2's score
and $\beta's$ denote the regression coefficients. SPSS computer programis used todetermine the regression coefficients, analyze the data and predict the exam scores based on the Test 1 and Test 2 scores.

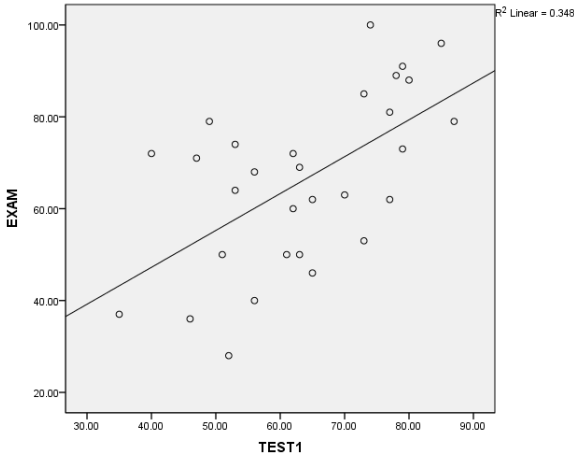
III. RESULT AND ANALYSIS

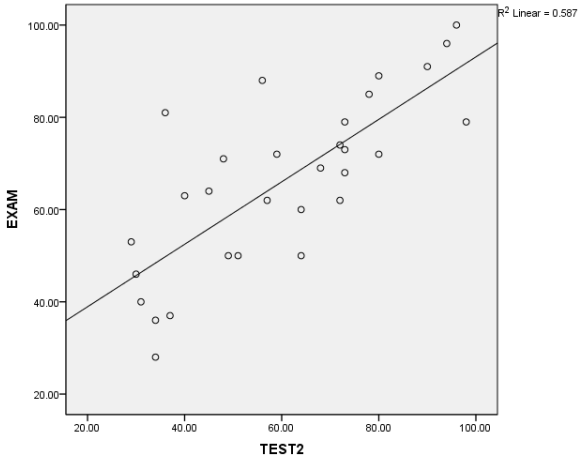
A multiple regression model need to satisfy the following four assumptions:

1. linearity,
2. normality,
3. reliability of measurement, and
4. homoscedasticity.

Linearity

Multiple regressioncan accurately estimate the relationship between independent variable (or variables) and dependent variable if their relationship is linear. Scatter graphcan be used to check the linearity between these variables. The two scatter graphs below showed that exists linear relation between each independent variable (Test 1 and Test 2) with the dependent variable (Final Exam).





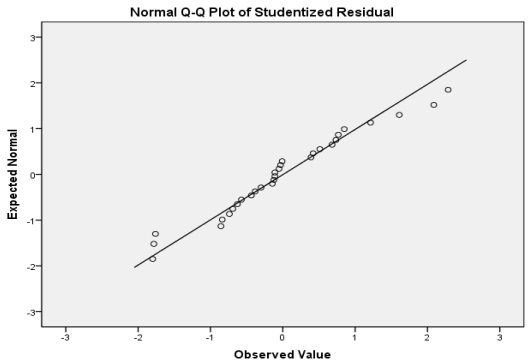
Normality

Regression model assumes that the variables must have normal distributions. The non-normally distributed variables can distort the relationship and the significance tests. Kolmogorov-Smirnov test and Shapiro-Wilk test can provide the inferential statistics on normality, as given in the table below. We can see the significant valuefor Kolmogorov-Smirnov Test is 0.142 > 0.05 and also for Shapiro-Wilk is 0.368 > 0.05. Therefore, we can conclude that for both tests, the variables have normal distributions. The variables are also been proven to have normal distributions because the points in the Q-Q plot form a relatively straight line.

Table 1: Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Studentized Residual	.139	30	.142	.963	30	.368

a. Lilliefors Significance Correction



Reliability of Measurement

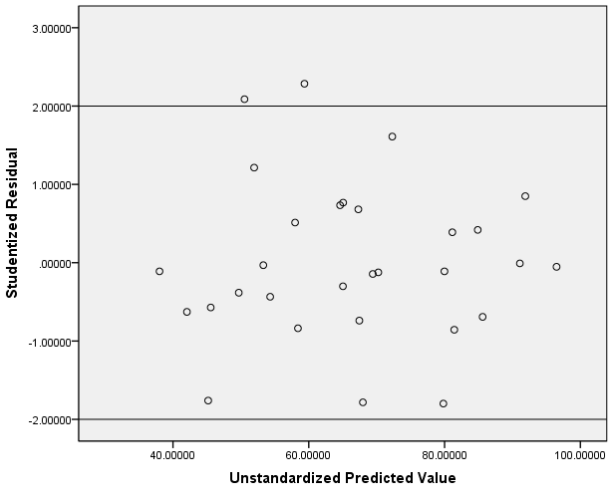
The multiple linear regression assumes that there is no multi-collinearity in the data. Multi-collinearity occurs when the independent variables are too highly correlated with each other. This can be checked using Pearson bivariate correlation. In the table below, we can see that correlation between Test 1 and Test 2 is 0.321 < 0.80. This satisfies the assumption of reliability of measurement.

Table 2: Correlations

		EXAM	TEST1	TEST2
Pearson Correlation	EXAM	1.000	.590	.766
	TEST1	.590	1.000	.321
	TEST2	.766	.321	1.000
Sig. (1-tailed)	EXAM	.	.000	.000
	TEST1	.000	.	.042
	TEST2	.000	.042	.
N	EXAM	30	30	30
	TEST1	30	30	30
	TEST2	30	30	30

Homoscedasticity

Homoscedasticity means that the variance of errors is the same for all independent variables. In the graphs below, we can see that the assumption for homoscedasticity is satisfied since almost all data are between 2 and -2.



Since all four assumptions have been fulfilled, the dependent and independent variables are now adequate for the next step in developing the regression model.

Regression model

Table 3: Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.848[a]	.719	.698	10.25770	1.896

a. Predictors: (Constant), TEST2, TEST1
b. Dependent Variable: EXAM

Table 4: ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7276.918	2	3638.459	34.579	.000[a]
	Residual	2840.949	27	105.220		
	Total	10117.867	29			

a. Predictors: (Constant), TEST2, TEST1
b. Dependent Variable: EXAM

Table 5: Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-1.272	9.443		-.135	.894	-20.647	18.103
TEST1	.522	.147	.383	3.561	.001	.221	.822
TEST2	.568	.095	.643	5.972	.000	.373	.763

a. Dependent Variable: EXAM

Table 3 shows that the multiple correlation coefficient (*R*) is 0.719. This value indicates that the independent variables (Test 1 and Test 2) have strong positive correlation with the dependent variable (Final Exam). The findings also showed that the relationship between Test 1 and Final Exam is 0.590 which is moderately positive. Whereas, the relationship between Test 2 and Final Exam is 0.766 which is highly positive correlation. The coefficient of determination (*R*-square) obtained is 0.666, meaning 66.6% of variation from the final exam can be explained by variations from Test 1 and Test 2. The remaining 33.4% is explained by other factors that are not in the model. From these results, we can conclude that our independent variables have strong predictive powers to the students' performance in Final Exam.ANOVA results in Table 4 reveals that the *p*-value is 0.000 < 0.05 which indicates that we shall reject H_0 and accept the alternative hypothesisH_1: students' performance in Final Exam is determined by their performance in Test 1 and Test 2. From the analysis in Table 5, the multiple linear regression model can be expressed as:

$$Y = -1.272 + 0.522X_1 + 0.568X_2$$

where*Y* is the predicted score in Final Exam, X_1 is the total score in Test 1, and X_2 is the total score in Test 2. The *p*-value of the estimated coefficients of X_1 is 0.001 < 0.05 and X_2 is 0.000 < 0.05, indicating that both X_1 and X_2 are significantly related to *Y*.InTable 6 below, we use the obtained regression model to find the predicted exam's scores and compare with the observed scores for 10 chosen students. We can see that the predicted and the observed exam scores have relatively small difference which indicates that the regression model is adequate to use.

Table 6: Comparison between the observed and the predicted score in final exam

Student	Test 1 (X_1)	Test 2 (X_2)	Observed Final exam	Predicted final exam (Y)	Difference predicted & observed
1	65	57	62	65.0	3.0
2	65	30	46	49.7	3.7
3	73	29	53	53.3	0.3
4	35	37	37	38.0	1.0
5	73.0	78.0	85.0	81.1	3.9
6	63.0	68.0	69.0	70.2	1.2
7	79.0	90.0	91.0	91.1	0.1
8	85.0	94.0	96.0	96.5	0.5
9	56.0	73.0	68.0	69.4	1.4
10	49.0	98.0	79.0	80.0	1.0

IV. CONCLUSION

The findings from this study has achieved the objective of developing a model that can predict the students' performance in final exam. The analysis has shown that the students who perform well in Test 1 and Test 2 have better chances of getting good scores in final exam, and vice versa. The value of *R*-square indicatez at least one of the predictor variables contributes to the prediction of the students' performance in final exam. Meanwhile, the rejection of H_0 indicates that there is significant relations between final exams' scores and the independent variables (Test 1 and Test 2).

The limitations of study are described below:

1. Small sample size.
2. Inconsistencies in choosing the right groups of students as sample.

For future research, we will consider developing the similar model for the current/future students who have different type of assessment and do the comparison to determine which assessment method gives better result in students' performance.

REFERENCES

1. Adeyemi, T.O., (2008) 'Predicting Students' Performance in Senior Secondary Certificate Examination from Performance in Junior Secondary Certificate Examinations in Ondo State, Nigeria', *Humanity & Social Sciences Journal* 3(1), pp. 26-36.
2. Adhatrao, K., Gaykar, A., Dhawan, A, Jha, R. and Honrao, V. (2013) 'Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms', *Computers and Society*, Cornell University Library.
3. Ayan, M. N. R and Garcia, M. T. C. (2008) 'Prediction of University Students' Academic Achievement by Linear and Logistic models', *The Spanish Journal of Psychology* 11, pp. 275-288.
4. Green, S. I. (2005) 'Student Assessment Precision in Mechanical Engineering Courses', *Journal of Engineering Education* 94, pp. 273-278.
5. Huang, S and Fang, N. (2010) 'Regression Models for Predicting Student Academic Performance In an Engineering Dynamics Course', *American Society for Engineering Education.*
6. Ibrahim, Z. and Rusli, D. (2007) 'Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression', *21st Annual SAS Malaysia Forum*, Kuala Lumpur.
7. Imbrie, P. K., Lin, J.J., Reid, K., and Malyscheff, A. (2008) 'Using Hybrid Data to Model Student Success in Engineering with Artificial Neural Networks', *Proceedings of the Research in Engineering Education Symposium*, Davos, Switzerland.
8. Karamazova, E., Zenku, T. J., and Trifunov, Z. (2017) 'Analysing and Comparing the Final Grade in Mathematics by Linear Regression Using Excel and SPSS', *International of Mathematics Trend and Technology (IJMTT)*, 52 (5), pp. 334-344.
9. Khan, W. S and Al-Zubaidy, S. (2017) 'Prediction of Student Performance in Academic and Military Learning Environment: Use of Multiple Linear Regression and Predictive Model and Hypothesis Testing', *International Journal of Higher Education*, 6 (4), pp. 152-160.
10. Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2010) 'Predicting Students' Performance in Distance Learning using Machine Learning Techniques', *Applied Artificial Intelligence*, 18(5), pp. 411-426.
11. Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., and Punch, W. F. (2003) 'Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System', *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conferences*, Boulder, CO.
12. Nghe, N. T., Janecek, P., and Haddawy, P. (2007) 'A Comparative Analysis of Techniques for Predicting Academic Performance', *Proceedings of the 37thASEE/IEEE Frontiers in Education Conferences*, Milwaukee, WI.
13. Thomas, E. H. and Galambos, N., (2004) 'What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analyis', *Research in Higher Education* 45, pp. 251-269.
14. Yousuf, A. (2000) 'Self-Efficacy and Vocational Interests in the Prediction of Academic Performance of Students in Engineering Technology', *Proceedings of the 2000 American Society for Engineering Education Annual Conference & Exposition*, St. Louis, MS.

AUTHORS PROFILE

Hazlina Darman Is working in School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

Sarah Musa Is working in School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

Rajasegeran Ramasamy Is working in School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.

Raja Rajeswari Is working in School of Actuarial Science, Mathematics, and Qualitative Study, Asia Pacific University of Technology and Innovation, Malaysia.