INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 12, Issue 3, (2023) E-ISSN: 2226-6348

Multiple Regression in Determining Affecting Factors Student Success in a Statistics Subject

Nur Syuhada Muhammat Pazil[1] and Norwaziah Mahmud[2]

[1]Mathematical Sciences Study, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Melaka Branch, Jasin Campus, 77300 Merlimau, Melaka, Malaysia, [2]Mathematical Sciences Study, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia

Corresponding Auhtor Email: norwaziah@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v12-i3/17410 DOI:10.6007/IJARPED/v12-i3/17410

Published Online: 20 September 2023

Abstract

This study aims to find the factor that affects students' success in Introductory Statistics Subjects based on a Multiple Linear Regression (MLR). Gender and assessment achievements such as test 1, test 2, quiz, assignment, group project, and final test marks were investigated as predictors. The dependent variable is the overall marks of the subject. The results shows that test 1, test 2, assignment, project, and final test have a significant difference to the overall marks of the statistics subject. This study was carried out using SPSS software. In order to determine the significant variables, further research can be done using more sample size and more variables.

Keywords: Academic Performance, Linear Regression, Statistics, Subject

Introduction

Increasing student behaviour in terms of attendance, assignment completion, performance in assessments, and class participation are the factors that contributes to university excellence. The quality of lecturers influences the quality of a university in ways such as strengthening teaching discipline, conducting continuous research, and providing high-quality educational services (Shedriko, 2021).

The multiple linear regression model (MLR) consists of one criterion, Y, also known as the response, predicted, outcome, or dependent variable, and predictors, p, also known as independent variables. There are numerous advantages to using a multiple regression model to analyse data. One of them is determining the relative influence of one or more predictor variables on the criterion value. Besides, it can identify outliers or anomalies. On the other hand, the relapse examination hypothesis can be exceptionally unappeasable since one excluded variable can make all relapse coefficients one-sided to an obscure degree and course (Klees, 2016). This demonstrates that any disadvantage of using a multiple regression model is usually due to the data used. MLR has numerous applications in nearly every field, including engineering, physical and chemical sciences, finance, administration, life, biological sciences, social sciences, and academics. The following are examples of MLR applications in academia.

Previous study by Yang et al (2018) used Multiple Linear Regression (MLR) a popular method in predicting student's academic performance. Mahmud et al (2022) employed Multiple Linear Regression (MLR) to investigate the factors that influence students' academic performance during the COVID-19 pandemic. It has been discovered that students' hometowns and the hours they spent preparing before class contributed significantly to the model. Tinuke Omolewa et al (2019) examine the student's performance using k-mean clustering and Multiple Linear Regression. It was found that, test score, quiz and assignment were the major factors in academic performance.

Furthermore, Weidlich and Bastiaens (2018) investigated a study on the impact of transactional distance on satisfaction in online distance learning. In this study, the dependent variable is satisfaction in online education. The independent variables are TD Student-Teacher (TDST), TD Student-Student (TDSS), TD Student-Content (TDSC), and TD Student-Technology (TDSTECH) as TD stands for transactional distance. This study revealed TDSTECH is the single most important independent variable or predictor of satisfaction in online distance learning for the chosen population. In addition, mediator analysis revealed that TDSTECH mediates the relationship between TD student-teacher and satisfaction, but not for TD student-content. However, there is no significant relationship between TD student-student and satisfaction.

Meanwhile, Rachmawati et al (2021) investigated the effect of online learning and parental guidance towards the result of XI social students' learning on Geography courses at SMAN 5 Jember. Independent variables in this study are online learning and parental guidance, and the dependent variable is the study result. This study concluded that online education and parental guidance could affect students' learning outcomes in Geography subject.

Hsu Wang (2019) studied the prediction of online behaviour and achievement by using self-directed learning awareness in flipped classrooms. The dependent variables in this study are the prediction of online behaviour and achievement, whereas the independent variable is self-directed learning factors. The results indicated three things: task value, intrinsic motivation, control of learning beliefs, and metacognition predict achievement. SRL awareness predicts online behaviours to a limited extent, and a combination of SRL awareness and online behaviours indicates that achievement is better than either one of the single-domain models.

Forson and Vuopala (2019) conducted online learning readiness of students enrolled in distance education in Ghana. The dependent variable is online learning readiness, whereas the independent variables are students' attitude, self-regulated learning, ICT skills, and collaborative skills. The study discovered that distance education students have positive attitude towards online learning. They also possess good self-regulated learning, cooperative and information communication and technology skills relevant for online learning.

From the previous studies, it is clear that linear regression can determine the relationship between two variables in education. Hence, this study aims to determine the factors that affect student success in statistics subject using Multiple Regression analysis.

Methodology

Secondary data was obtained from all students who took the subject introduction to statistics at UiTM Cawangan Melaka Kampus Jasin in the two semesters of 2021. There were 236 undergraduate students in the Faculty of Plantation and Agrotechnology who took that subject. Test 1, test 2, quiz, assignment, group project, final test, gender and overall marks are the input parameters used.

The data analysis method employs statistical and logical techniques to describe, illustrate, and evaluate data. It began with the validation of assumptions and the evaluation of the fitted model. Assumption must be met during the validating assumptions stage, or the process must be restarted from the beginning. Throughout the evaluation of the fitted model, the estimated model was tested three times before it could be claimed as the best model and used to forecast values. Data was analysed using SPSS to represent the findings.

Findings and Discussions

a) Set of Variables

The study analyzed the factor on students' performance in subject by measuring students' marks on seven factors which are test1, test 2, quiz, assignment, group project, final test, gender and overall mark in Table 1.

Table 1
Set of variables

Variable	Type of variable	Variable Code
Test 1	Quantitative continuous	Test1
Test 2	Quantitative continuous	Test2
Quiz	Quantitative continuous	Quiz
Assignment	Quantitative continuous	Assignment
Group Project	Quantitative continuous	Project
Final test	Quantitative continuous	FinalTest
Overall mark	Quantitative continuous	OverallMarks
Gender	Qualitative	Gender

b) Forming the Model

$$y_i = B_0 + B_1X_{i1} + B_2X_{i2} + B_3X_{i3} + B_4X_{i4} + B_5X_{i5} + B_6X_{i6} + B_7X_{i7} + e_i$$

Where

y_i : OverallMarks

B_0 : the intercept

B_1.....B_7 : The regression coefficient for independent variables

X_{i1} : Test1

X_{i2} : Test2

X_{i3} : Quiz

X_{i4} : Assignment

X_{i5} : Project

X_{i6} : FinalTest

X_{i7} : Gender

e_i : Model's error term or residuals

Table 2
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	0.937[a]	0.878	0.874	5.08642	1.562

a. Predictors: (Constant), Gender,Quiz, FinalTest, Project,Test1,Assignment,Test2
b. Dependent Variable: OverallMarks

Table 3
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42450.619	7	6064.374	234.402	0.000[b]
	Residual	5898.748	228	25.872		
	Total	48349.367	235			

a. Dependent Variable: OverallMarks
b. Predictors: (Constant), Gender,Quiz, FinalTest, Project,Test1,Assignment,Test2

Table 4
Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	9.104	2.739		3.323	.001					
	Test1	.134	.023	.230	5.769	.000	.435	.357	.133	.335	2.983
	Test2	.064	.018	.149	3.516	.001	.388	.227	.081	.299	3.344
	Quiz	-.032	.017	-.084	-1.856	.065	.322	-.122	-.043	.260	3.841
	Assignment	-.033	.016	-.083	-1.991	.048	.340	-.131	-.046	.305	3.281
	Project	.074	.028	.068	2.620	.009	.302	.171	.061	.794	1.259
	FinalTest	.703	.020	.836	34.677	.000	.903	.917	.802	.921	1.085
	Gender	-.027	.697	-.001	-.038	.970	-.109	-.003	-.001	.923	1.084

a. Dependent Variable: OverallMarks

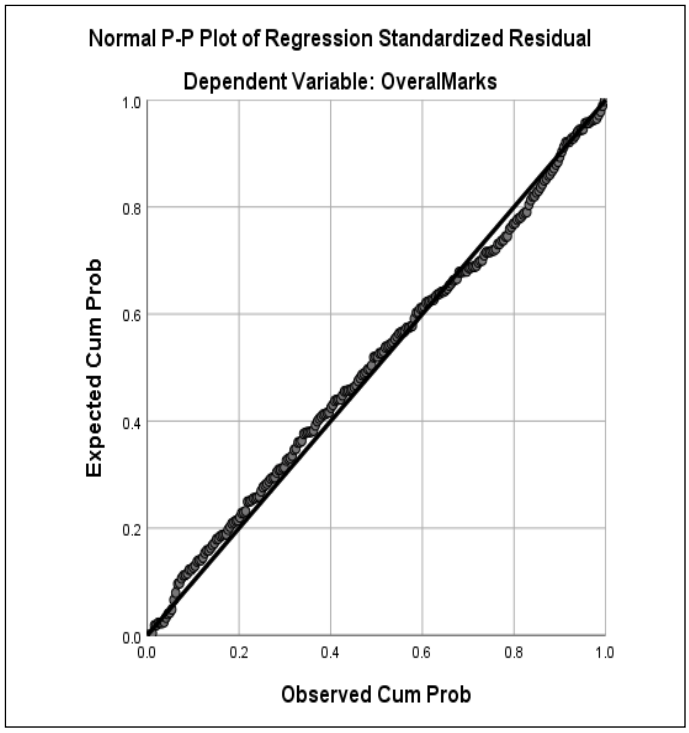

Figure 1: P-P plot

Validating Assumption

In regression analysis, many assumptions about the model and the Multiple Linear Regression (MLR) model are one of the fussier of the statistical techniques as it makes several assumptions about the data. If one or more assumptions are violated, then the model in hand is no longer reliable and not acceptable in estimating the population parameters (Daoud, 2018). In this study, four assumptions were discussed.

a) The relationship between independent variables and dependent variables is linear

Based on Table 4, it shows that gender has a weak negative correlation with overall marks (r -0.109). Test 1, test 2, quiz assignment and project have weak positive correlations with overall marks. The correlation between the two variables above shows that no factor is controlled or held constant. Final test has a strong positive correlation with overall marks (r = 0.903). The correlation between the two variables above shows that no factor is controlled or held constant.

b) Checking Multicollinearity

Multicollinearity is when there is a correlation between independent variables in a model. Based on Table 4, there is no multicollinearity as the value of VIF scores are below 10 and the tolerance scores are above 0.1. Therefore, there is no absence of multicollinearity among the independent variables.

c) The Values of the Residual are Independent

The Durbin-Watson statistic in Table 2 shows that this assumption had been met, as the obtained value is close to 2, which is 1.562

d) The values of Residuals are normally distributed

The P-P plot (Figure 2) for the model suggested that the assumption of normality of the residuals may have been violated. However, as only extreme deviations from normality are likely to significantly impact the findings, the result was probably still valid.

Evaluate the Model

Evaluating the estimated model is a necessary but often overlooked procedure. However, it is a necessary prerequisite before the estimated model can be claimed as the best model and used to forecast values. The subsections that follow describe some of the most common statistical testing procedures.

a) Fitness of the Model

This is a test for the overall fitness of the model. The examination will reveal whether all or part of the independent variables should remain in the model. The test criterion used is the *F-test statistic*. The null hypothesis to be tested states that all coefficients in the model are equal zero, that is

H_0 : The regression model is not significant

H_1 : The regression model is significant

The overall F-test can be found in the ANOVA table in the statistical output. To interpret the F-test of significance, the *p*-value for the F-test must be compared to a 5% significance level. If the *p*-value is less than the significance level chosen, the data provide sufficient evidence to conclude that the independent variables in the model improve the fit.

This means that none of the independent variables provides any information for fitting the model, and hence the model is rejected. From Table 3, the *p*-value 0.000 is less than the significance level of 0.05. The data provide sufficient evidence to conclude that the regression model is significant.

b) Goodness of Fit

The standard measure of the goodness of fit is the *coefficient of determination*, R^2. From Table 2, the coefficient of determination, R^2 shows that 87.8% of the total variation in overall marks can be explained by gender, quiz, final test, project, test1, assignment, test2 while the others 12.2% are cause by errors. Therefore, the is quite good (Dhakal, 2018).

c) Statistical Significance of the Independent Variables

The independent is significant when the *p-value* is less than 0.05. from Test 1 (B = 0.134, p =0.000 < 0.05), test 2 (B = 0.064, p =0.001 < 0.05), assignment (B = -0.033, p =0.048 < 0.05), project (B = 0.074, p =0.009 < 0.05) and final test (B = 0.703, p=0.000 < 0.05) contributed significantly to the model while quiz (B = -0.032, p =0.065 > 0.05) and gender (B = -0.027, p=0.970 > 0.05 did not. These values are presented in Table 4. From this value, it can be concluded that test 1, test2, assignment, project and final test are significant variables towards overall marks.

Estimated Model Coefficient

From Table 4, the estimated model coefficient is:

OverallMarks = 9.104 + 0.134 Test1 + 0.064 Test2 -0.032 Quiz – 0.033 Assignment +0.074 Project + 0.703 FinalTest – 0.027 Gender

The analysis revealed that the factors affecting students' success in statistics subject are test 1, test 2, assignment, project and final test.

Conclusion and Recommendations

To achieve the objective of the study, which is to determine the factors that affect students' success in statistics subject, test 1, test 2, quiz, assignment, group project, final assessments, and gender were analysed using multiple linear regression. From the results, it is shown that the variables that are significant are test 1, test 2, assignment, project and final test. Therefore, these variables are more important to success in a statistics subject. Similarly, Syuhada et al (2023) discovered that test 2 and final test are significant variables in determining the factor affect in statistics subject. There are several gaps in knowledge and limitation while investigating the factors that affect students' success in statistics subject. For future research, it is recommended to have more variables and collecting data and add more sample size to gain more accurate results.

References

Daoud, J. I. (2018). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, *949*(1). https://doi.org/10.1088/1742-6596/949/1/012009

Dhakal, C. P. (2018). Multiple Linear Regression in SPSS. *Article in International Journal of Science and Research*. https://www.researchgate.net/publication/333973273

Forson, I. K., & Vuopala, E. (2019). Online learning readiness: perspective of students enrolled in distance education in Ghana. *The Online Journal of Distance Education and E-Learning*, *7*(4), 277–294. www.tojdel.net

Hsu Wang, F. (2019). On prediction of online behaviors and achievement using self-regulated learning awareness in flipped classrooms. *International Journal of Information and Education Technology*, *9*(12), 874–879. https://doi.org/10.18178/ijiet.2019.9.12.1320

Klees, S. J. (2016). Inferences from regression analysis: are they valid? *Real-World Economics Review*, *74*, 85–97.

Mahmud, N., Pazil, M. N. S., & Azman, N. A. N. (2022). The significant factors affecting students' academic performance in online class: Multiple linear regression approach. *Jurnal Intelek*, *17*(2), 1–11. https://doi.org/10.24191/ji.v17i2.17896

Rachmawati, S., Mutrofin, & Sumardi. (2021). The effect of online learning and parental guidance towards the result of XI social students' learning on Geography course at SMAN 5 Jember. *IOP Conference Series: Earth and Environmental Science*, *747*(1). https://doi.org/10.1088/1755-1315/747/1/012028

Shedriko. (2021). Binary logistic regression in determining affecting factors student graduation in a subject. *Jurnal Teknologi Dan Open Source*, *4*(1), 114–120. https://doi.org/10.36378/jtos.v4i1.1401

Syuhada, N., Pazil, M., Mahmud, N., Baharom, N., & Hafawati, S. (2023). Logistic regression in determining affecting factors student success in an introductory statistics subject. *Jurnal Intelek*, *18*(1), 9–16. https://doi.org/https://doi.org/10.24191/ji.v18i1.20133

Omolewa, T. O., Oladele, T. A., Adeyinka, A., & Oluwaseun, R. O. (2019). Prediction of student's academic performance using k-eans clustering and multiple linear regressions. *Journal of Engineering and Applied Sciences*, *14*(22), 8254–8260.

https://doi.org/10.36478/jeasci.2019.8254.8260

Weidlich, J., & Bastiaens, T. J. (2018). Technology matters - the impact of transactional distance on satisfaction in online distance learning. *International Review of Research in Open and Distance Learning*, *19*(3), 222–242. https://doi.org/10.19173/irrodl.v19i3.3417

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, *26*, 170–176. https://doi.org/10.2197/ipsjjip.26.170