International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Forecasting Academic Performance using Multiple Linear Regression

Yee Mon Khaing, Aung Cho

University of Computer Studies, Maubin, Myanmar

How to cite this paper: Yee Mon Khaing | Aung Cho "Forecasting Academic Performance using Multiple Linear Regression" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1187-1189, **https://doi.org/10.31142/ijtsrd26517**

IJTSRD26517

ABSTRACT

Regression is one of the most powerful statistical methods used in educational researches. This paper shows the important instance of regression methodology called Multiple Linear Regression (MLR) and proposes a framework of the forecasting of the students' test scores, based on Intelligence Quotient (IQ) and the number of hours that the students studied. This paper was applied the aid of the Statistical Package for Social Sciences (SPSS) version 23 and PYTHON version 3.7.

KEYWORDS: *Multiple Linear Regressions (MLR), Statistical Package for Social Sciences (SPSS)*

1. INTRODUCTION

Education is important for the personal, social and economic development of the nation. This paper is useful for students and teachers to improve academic performance. The main purpose of this analysis is to know to what extent is the students' test scores influenced by the two independent variables, IQ and study hours. It used multiple linear regression method for data forecasting and ANOVA algorithm for data significant.

2. SPSS

The "Statistical Package for the Social Sciences" (SPSS) is a package programs for manipulating, analyzing, and presenting data. SPSS is widely used by market researchers, health researchers, survey companies, government entities, education researchers, marketing organizations, data miners, and many more for the processing and analyzing of survey data. [2]

3. Regression

Regression is a powerful statistical method used in education, finance, investing and other disciplines that allow estimating the relationships between one dependent variable and one or more independent. As with most statistical analyses, the goal of regression is to summarize observed data as simply, usefully, and elegantly as possible. The two basic types of regression are *simple linear regression* and *multiple linear regression*, although there are *non-linear regression* methods for more complicated data and analysis. Simple linear regression uses one independent variable to predict the outcome of the dependent variable whereas multiple linear regression uses two or more independent variables to predict the outcome of the dependent variable. [4]

4. Methodology [3][5]

4.1. Multiple Linear Regression

Multiple linear regression (MLR), known as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.

$$y = mx1 + mx2 + mx3 + b$$

where,
y = the dependent variable of the regression equation
m = slope of the regression equation
x1= first independent variable of the equation
x2 = second independent variable of the equation
x3 = third independent variable of the equation
b=constant of the equation

4.2. R Square

R-Squared is a statistical measurement in a regression that calculates the proportion of variance in a dependent variable that is explained by an independent variable or variables. R-squared tells how well the data fit the regression model (the goodness of fit). R-squared can take any values between 0 and 1. R-squared is better if the values are closer to 1.

$$R\ Square\ Formula = r^2$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum xy)^2]}}$$

where,
r = the correlation coefficient
n = number in the given dataset
x = first variable in the context
y = second variable

4.3. ANOVA Table

ANOVA is the short form of analysis of variance. ANOVA is a statistical tool which is generally used on random variables.

It involves group not directly related to each other in order to fine whether exist any common means.

4.4. Significance F-Value

F-Test is any test that uses F-distribution. F value is a value on the F distribution. Various statistical tests generate an F value. The value can be used to determine whether the test is statistically significant. In order to compare two variances, one has to calculate the ratio of the two variances:

$$F = \sigma_1^2/\sigma_2^2$$

where,

σ_1^2 = larger sample variance

σ_2^2 = smaller sample variance

4.5. P-Value

P is a statistical measure that helps researchers to determine whether their hypothesis is correct. It helps determine the significance of result. P-Value is a number between 0 and 1. Calculating P-Value from a Z Statistic

statistic z

$$Z = \frac{\hat{p} - p0}{\sqrt{\frac{p0(1-p0)}{n}}}$$

where,

p̂ is Sample Proportion

p0 is assumed Population Proportion in the Null Hypothesis

n is the Sample Size

5. Testing

Table-1: Sample Data

Test Score	IQ	Study Hours
100	125	30
95	104	40
92	110	25
90	105	20
85	100	20
80	100	20
78	95	15
75	95	10
72	85	0
65	90	5

The table provides us the data needed to perform the multiple regression analysis. We can predict that there is a relation between Test Score (Output) and IQ and Study Hours (Input).

Table-2: Regression Values

Summary Output

Regression Statistics	
Multiple R	0.951
R Square	**0.905**
Adjusted R Square	0.878
Standard Error	3.875
Observations	10.000

R-squared is better if the values are closer to 1. In the table, the result of R Square is 0.905 that's good. Therefore, *the proportion of the variance* is 91% for Test Score that is explained by IQ and hours spent in study.

Table-3 and Table-4: ANOVA Table

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1004.484	502.242	**33.446**	**0.00026**
Residual	7	105.116	15.017		
Total	9	1109.600			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	23.156	15.967	1.450	0.190	-14.600	60.913	-14.600	60.913
IQ	0.509	0.181	2.818	**0.026**	0.082	0.937	0.082	0.937
Study Hours	0.467	0.172	2.717	**0.030**	0.061	0.874	0.061	0.874

As Table 3, Significance F and P-values

This table tests the statistical significance of the independent variables as predictors of the dependent variable. The last column of the table shows the results of an overall F test. The F statistic (33.4) is big, and the p value (0.00026) is small. This indicates that one or both independent variables have explanatory power beyond what would be expected by chance.

As table-4, Significance of Regression Coefficients

The coefficients table shows the following information each coefficient: its value, its standard error, a t-statistic, and the significance of the t-statistic. In this table, the t-statistics for IQ and study hours are both statistically significant at the 0.05 level. This means that IQ contributes significantly to the regression after effects of study hours are taken into account. And study hours contribute significantly to the regression after effects of IQ are taken into account.

Table-5: RESIDUAL OUTPUT

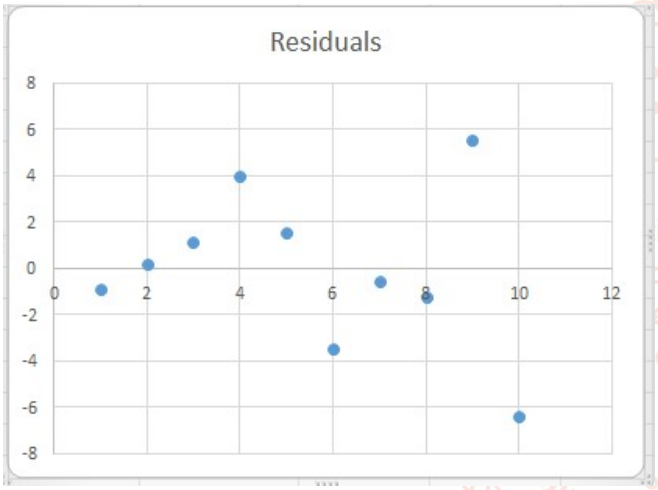
The result of coefficients can use to do a forecast. The regression line is: y = Test Score = 23.156 + 0.509 * IQ + 0.467 * Study Hours. In other words, for each unit increase

in IQ, Test Score increase with 0.509 units. For each unit increase in Study Hours, Test Score increases with 0.467 units.

RESIDUAL OUTPUT

Observation	Predicted Test Score	Residuals
1	101	-0.849
2	95	0.177
3	91	1.128
4	86	4.011
5	83	1.558
6	83	-3.442
7	79	-0.559
8	76	-1.224
9	66	5.542
10	71	-6.341

Graph-1: Residuals Values



(C) Analytical Views

As table-5 and graph-1, Residuals

The residuals show you how far away the actual data points are from the predicted data points (using the equation). For example, the first data point equals 100. Using the equation, the predicted data point equals 23.156 + 0.509 * 125 + 0.467 * 30 = 100.894, giving a residual of 100 – 100.894 = -0.894.

(D) Other Analytical Result

As the result of table-6, will see if there is higher students' IQ, higher students' scores. Therefore, the educational leaders need to aware the students' IQ and should be more careful and teach if the students need to learn the lessons.

As the result of table-7, will see if there is more hours spent in study, higher students' scores. The exam performances of the students that more hours spent in study higher than other. So, the students will need to spend more hours to study the lessons.

The exam performances of the students are communicated to their IQ and hours spent in study.

Table-6

IQ	80	90	100	110
Study Hour	20	20	20	20
Predicted Test Score	73.254	78.348	83.442	88.537

Table-7

IQ	100	100	100	100
Study Hour	10	20	30	40
Predicted Test Score	78.771	83.442	88.114	92.785

6. Conclusion

SPSS data analysis tools are valuable in education, business and marketing fields. It is very good for presentation report by graphical design. This paper is useful for students and teachers to improve academic performance and regression is one of the most powerful statistical methods used in educational researches. This paper shows the important instance of regression methodology called Multiple Linear Regression (MLR) and proposes a framework of the forecasting of the students' test score, based on Intelligence Quotient (IQ) and the number of hours that the students studied by using SPSS software.

References

[1] IBM SPSS Statistics 24 Algorithms pdf book
[book style]

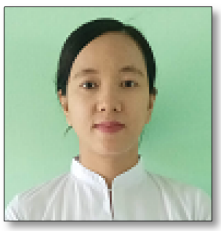
[2] A handbook of statistical analyses using SPSS / Sabine, Landau, Brian S. Everitt, ISBN 1-58488-369-3
[book style]

[3] https://www.exceleasy.com

[4] https://www.investopedia.com/terms/r/regression.asp

[5] https://www.wallstreetmojo.com

Author Profile



Yee Mon Khaing received the M.C.Sc. (Computer Science) degree from Univerrsity of Computer Studies, Yangon in 2009. Now, I served as a teacher at the Application Department, University of Computer Studies, Maubin, Myanmar.

Aung Cho received the B.A.(Eco) degree from Yangon University in 1987 and M.I.Sc.(Information Science) degree from University of Computer Studies, Yangon in 2001. After got Master degree, I served as a teacher at the software, information science and application departments of the computer universities. I am now with University of Computer Studies, Maubin.