Statistical analysis of seventh grade mathematics grades

Leonard Sheehy

Statistics for Educational Research

New Jersey City University

Part 1: Introduction

Preparing students for careers in science, technology, engineering and mathematics (STEM) is essential not only for their success, but for the continued high achievement of our country as we compete in a growing global economy. The President's Council of Advisors on Science and Technology stated STEM education will determine whether the United States will remain a leader among nations and whether we will be able to solve immense challenges in such areas as energy, health, environmental protection, and national security (https://www.whitehouse.gov). This paper will analyze existing data in an attempt to predict second marking period mathematics averages based on factors that influence a student's ability to perform well in that subject. Data was collected from Talisay City State College for the purpose of this assignment. The data set consists of forty nine students with one missing (Alcantara, 2014). The formal hypotheses are:

$$H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H\alpha : \text{At least one } \beta \neq 0$$

$$\alpha = 0.05$$

A multiple regression study will be implemented in an attempt to predict student marking period averages. T-test and ANOVA were not chosen because those tests do not predict values.

Data Description

The data includes information from which two categorical and two quantitative values were selected. Seventh grade student's second marking period average was chosen as the response value. This value is an average calculated by the student's teacher and provided to parents via electronic report card. After school activities and mathematics level were chosen as

categorical value. The data set has been populated with a zero if the student does not participate in after school activities and a one if they do participate in after school activities. Students in standard level mathematics class have a zero in the math level category while students in the advanced level have a one. The number of hours students spend doing homework and first test score were chosen as the quantitative values.

The number of students enrolled or not enrolled in after school activities is displayed in table 1. Twenty students are enrolled in afterschool activities while twenty eight were not. This variable was chosen to determine if the time spent in these activities has a negative effective on second marking period grades because students are not focusing on homework or studying while engaged in other undertakings. Some believe that increased time on after school activities such as sports or art enrichment assist in keeping students stimulated and requires them to schedule their time more efficiently, thereby increasing their level of achievement.

After school activity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not enrolled	28	59.2	59.2	59.2
	Enrolled	20	40.8	40.8	100.0
	Total	48	100.0	100.0	

Table 1 number of students in after school activities

The number of children enrolled in standard and advanced level mathematics classes is listed in table 2. Twenty two students were enrolled in standard level mathematics while twenty six were not. This variable was chosen to determine if the level influences second marking period average. Students are thought to score higher when placed in classes with students of

their skill level. An alternative belief is that higher achieving students act as role models and can assist other students with understanding concepts.

Math level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Stand	22	44.9	44.9	44.9
	Adv	26	55.1	55.1	100.0
	Total	48	100.0	100.0	

Table 2 number of students in standard and advance level mathematics

Table 3 shows the descriptive statistics with N=48. The minimum for the second period average was 71, zero for hours doing homework, 37 for test one. The maximum for the second period average was 98, 5.0 for hours doing homework, and 96 for test one. The mean for the second period average was 88.6, 1.87 for hours doing homework, and 79.41 for test one. The standard deviation for the second period average was 6.5, 1.15 for hours doing homework, and 11.78 for test one. Hours doing homework was selected to determine if second marking period average increases with more time allotted to homework. First test was chosen to decide if a child's success rate increases for the second marking period average or if their level of achievement remains unchanged. This can have implications on motivational factors of children in mathematics classes because if they feel they can improve as the marking period goes on, they will be more likely to apply themselves at a higher level.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
2nd mp average	48	71	98	88.67	6.502

Hours doing hw	48	0	5	1.87	1.145
Test 1	48	37	96	79.41	11.799
Valid N (listwise)	48				

Table 3 descriptive statistics of seventh grade mathematics scores of quantitative values

Outliers

Consideration for the existence of outliers is prudent at this time. Review of illustration 1 shows that no subjects are determined to be outliers because the standardized residual of greater than 3.3 or less than -3.3 did not exist.

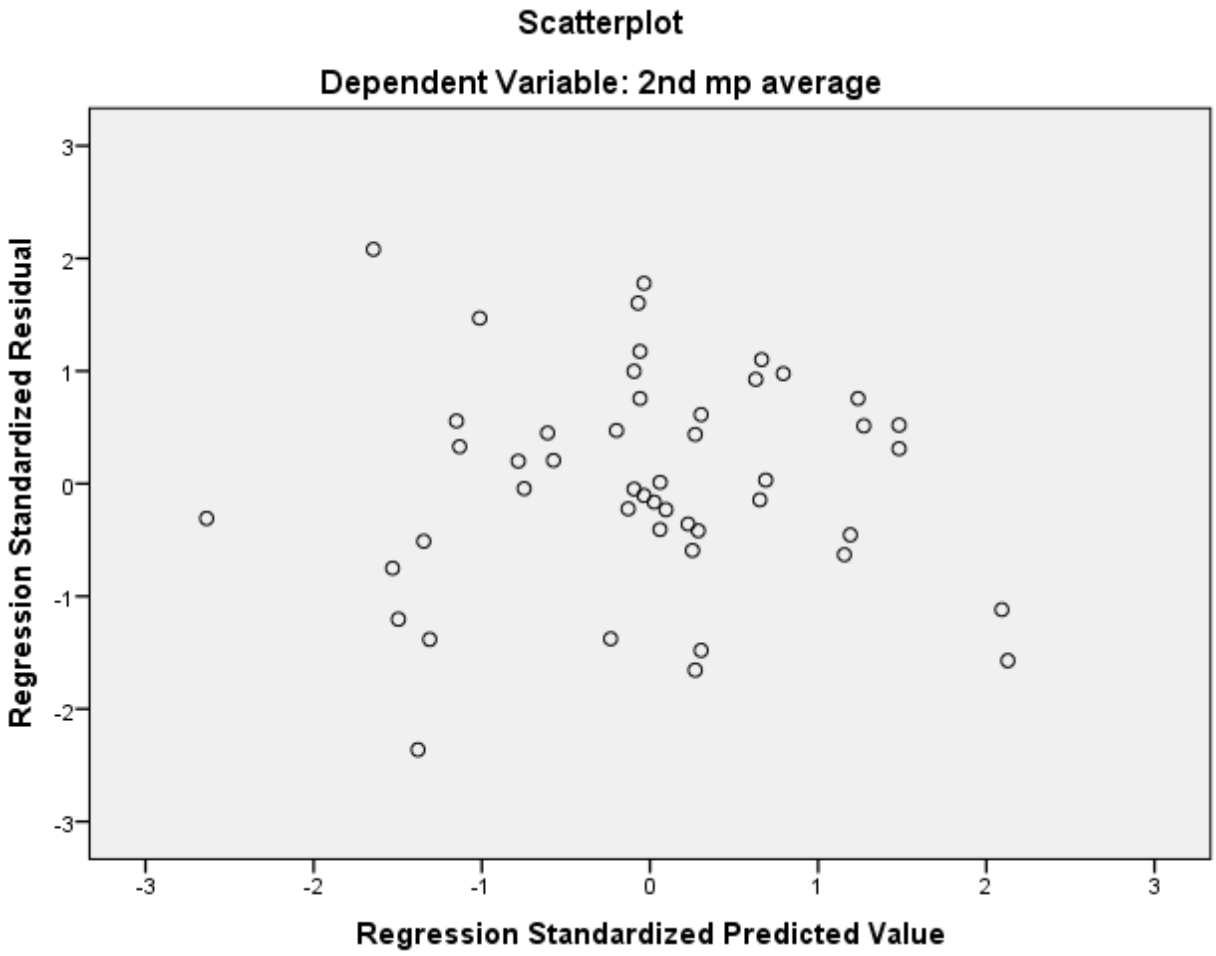

Illustration 1 Scatterplot of dependent variable.

Additional measurement of outliers was conducted by reviewing the Mahalanobis distance. Table 4 shows that maximum value of 16.87 does not exceed the critical level of 18.72. Therefore no outliers exist.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	76.47	98.51	88.67	4.624	49
Std. Predicted Value	-2.637	2.128	.000	1.000	49
Standard Error of Predicted Value	1.074	2.945	1.499	.366	49
Adjusted Predicted Value	76.98	100.60	88.60	4.803	48
Residual	-11.284	9.940	.073	4.563	48
Std. Residual	-2.362	2.080	.015	.955	48
Stud. Residual	-2.527	2.235	.015	1.016	48
Deleted Residual	-12.923	11.474	.069	5.170	48
Stud. Deleted Residual	-2.707	2.350	.013	1.040	48
Mahal. Distance	1.394	16.874	3.918	2.792	49
Cook's Distance	.000	.186	.027	.045	48
Centered Leverage Value	.030	.359	.083	.059	49

a. Dependent Variable: 2nd mp average

Table 4 Residual Statistics

The Model Hypothesis

The variables discussed above will be applied to a multiple regression model to predict the marking period scores of a student given the input variables discussed above. Testing each beta for significance, with the null hypothesis that such a predicative model does not exist (H_o: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) and the alternative hypothesis that at least one variable's beta coefficient is not zero (H_α: at least one $\beta \neq 0$). During this analysis the alpha level of 0.05 ($\alpha = 0.05$) will be applied to test statistical strength.

Model 1

Correlations

The Pearson's R for each of the variables considered is listed below in Table 4. The Pearson's R values demonstrate the following relationships when comparing to the second marking period average:

- After school activity ($r = -0.035$, $p < 0.05$) weak, negative and statistically significant.
- Math level ($r = 0.453$, $p > 0.05$) positive, strong and statistically insignificant relationship.
- Test 1 ($r = 0.447$, $p < 0.05$) strong, positive and statistically significant relationship.
- Hours doing homework ($r = 0.614$, $p < 0.05$) strong, positive and statistically significant relationship.

A correlation matrix of the variables, demonstrated in table 5 shows the relationships between the variables. Reviewing how each independent variable correlates with the dependent variable (second marking period average) shows that hours doing homework is the strongest followed by math level and test one. After school activities had an alpha value greater than 0.05 ($p = .253$).

Correlations

		2nd mp average	After school activity	Math level	Test 1	Hours doing hw
Pearson Correlation	2nd mp average	1.000	-.035	.453	.447	.614
	After school activity	-.035	1.000	-.169	-.175	.097
	Math level	.453	-.169	1.000	.239	.401
	Test 1	.447	-.175	.239	1.000	.232
	Hours doing hw	.614	.097	.401	.232	1.000

Sig. (1-tailed)	2nd mp average	.	.407	.001	.001	.000
	After school activity	.407	.	.123	.115	.253
	Math level	.001	.123	.	.049	.002
	Test 1	.001	.115	.049	.	.055
	Hours doing hw	.000	.253	.002	.055	.
N	2nd mp average	48	48	48	48	48
	After school activity	48	48	48	48	48
	Math level	48	48	48	48	48
	Test 1	48	48	48	48	48
	Hours doing hw	48	48	48	48	48

Table 5 Correlation of variables used in multiple regression.

Negative relationships (one value increases while the other decreases) were noted between math level and after school activity (r = - 0.169), and also with after school activity and test one (r = - 0.175), shown in table 4. They both had alpha values greater than 0.05 (p = 0.123 and p = 0.115 respectively). Math level and hours doing homework also had a noteworthy correlation (r = 0.401) with an alpha level p = .002.

Model 1

Initial Multiple Regression and Prediction

Tables 6 demonstrates the model summary. The adjusted R-square value of 0.460 indicates that 46% of the level of second quarter math average comes can be explained by at least one of the predictor values. The adjusted R-square was used due to the low N value.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.711[a]	.506	.460	4.778

a. Predictors: (Constant), Hours doing hw, After school activity, Test 1, Math level

b. Dependent Variable: 2nd mp average

Table 6 Model summary

The multiple regression to predict second quarter average from after school activities, math level, test one and hours doing homework was a statistically significant $F(4,43)=11.00$, $p < .05$, $R^2 = 0.506$ as show in table 7. It demonstrates a good model allowing rejection of the null hypothesis and accepts that at least one model coefficient is a significant predictor of second quarter average. A significate p-value of less than 0.05 is shown in table 7.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1005.121	4	251.280	11.008	.000[b]
	Residual	981.546	43	22.827		
	Total	1986.667	47			

a. Dependent Variable: 2nd mp average

b. Predictors: (Constant), Hours doing hw, After school activity, Test 1, Math level

Table 7 ANOVA

The Coefficients

Focus is directed on the individual coefficients to decide if there is statistical significance and to determine the effect of each variable on the model. Table 8 shows after school activity and math level both have p-values greater than 0.05 (after school activity p = .972, math level p = .133) and are contenders for exclusion from the model. Hours doing homework has the highest have statistical significance.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	69.482	4.951		14.033	.000	59.496	79.468					

After school activity	.051	1.470	.004	.035	.972	-2.914	3.016	-.035	.005	.004	.911	1.098
Math level	2.539	1.567	.196	1.620	.113	-.621	5.698	.453	.240	.174	.783	1.277
Test 1	.161	.062	.293	2.583	.013	.035	.287	.447	.366	.277	.893	1.119
Hours doing hw	2.650	.688	.467	3.853	.000	1.263	4.037	.614	.507	.413	.784	1.276

a. Dependent Variable: 2nd mp average

Table 8 Coefficients

Model 2

Even though the first model seemed to be adequate based on the p-value and F-ratio, I decided to attempt to improve the outcome. The second model consisted of two variables: hours doing homework and test one score. The data set remained the same for model 2 and can be referred to in table 4 for review of descriptive statistics.

Model Summary

The model 2 summary did not demonstrate improvements against model 1 in reference to the R-squared value as revealed in table 9.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.689[a]	.475	.451	4.815

a. Predictors: (Constant), Hours doing hw, Test 1

b. Dependent Variable: 2nd mp average

Table 9 Model 2 summary

The adjusted R-square value decreased to 0.451 meaning that that 45% of the level of second quarter math score can be explained by at least one of the predictor values. The multiple regression to predict second quarter average from test one and hours doing homework was a

statistically significant $F(2,45)=20.34$, $p < .05$, $R^2 = 0.451$ as show in table 10. It demonstrates a good model allowing rejection of the null hypothesis and accepts that at least one model coefficient is a significant predictor of second quarter average. A significate p-value of less than 0.05 is shown in table 10.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	943.182	2	471.591	20.337	.000[b]
	Residual	1043.485	45	23.189		
	Total	1986.667	47			

a. Dependent Variable: 2nd mp average

b. Predictors: (Constant), Hours doing hw, Test 1

Table 10 Model 2 ANOVA

Coefficients

As referenced in table 11 the p-value for test one decreased to 0.006. There is a positive correlation between test one and second quarter score. Also, there is a positive correlation between hours doing homework and second quarter score.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Tolerance
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	
1	(Constant)	68.846	4.779		14.407	.000	59.222	78.471				
	Test 1	.178	.061	.322	2.902	.006	.054	.301	.447	.397	.314	.946

Hours doing hw	3.061	.631	.539	4.852	.000	1.790	4.331	.614	.586	.524	.946

a. Dependent Variable: 2nd mp average

Table 11 Model 2 Coefficients

The standardized beta reveals that hours doing homework $\beta = 0.539$ has a greater effect on second quarter score than test one $\beta = 0.322$.

The Regression Model

The model equation can be constructed because suitable variables with significance are in place.

$$\text{Second quarter average} = b_0 + (b_1 \times \text{test one}) + (b_2 \times \text{hours doing homework})$$

$$\text{Second quarter average} = 68.84 + (0.178 \times \text{test one}) + (3.06 \times \text{hours doing homework})$$

Table 12 shows the second quarter averages produced by the model along with actual second quarter averages and residual values. The residual value is determined by calculating the difference between the observed value of the dependent variable and the predicted value. It assists in determining the accuracy of the model. Each data point has one residual. The model predicted case 43 and 44 would be higher than the actual value with substantial residuals. The model predicted cases 3, 4, 5, 6, 15, 30, 32, 38, 47, and 48 with residuals less than 1.

Case	Test one	Hours	Actual 2nd Qtr Avg	Predicted 2nd Qtr Avg	Residual
1	55	1	91	82.69	8.31
2	57	2	91	84.046	6.954
3	70	1	85	85.36	-0.36
4	70	3	88	87.36	0.64
5	70	3	88	87.36	0.64
6	72	2	87	86.716	0.284
7	75	0	78	85.25	-7.25
8	78	2	82	87.784	-5.784
9	79	2	93	87.962	5.038
10	43	0	75	79.554	-4.554
11	80	0	80	86.14	-6.14
12	80	1	86	87.14	-1.14
13	80	2	90	88.14	1.86
14	81	5	93	91.318	1.682
15	84	1	88	87.852	0.148
16	84	2	96	88.852	7.148
17	37	2	86	80.486	5.514
18	85	2	88	89.03	-1.03
19	85	2	88	89.03	-1.03
20	85	1	92	88.03	3.97
21	85	1	94	88.03	5.97
22	86	1	87	88.208	-1.208
23	88	3	91	90.564	0.436
24	88	3	92	90.564	1.436
25	88	2	96	89.564	6.436
26	90	3	97	90.92	6.08
28	96	3	98	91.988	6.012
29	84	1	93	87.852	5.148
30	85	1	88	88.03	-0.03
31	89	3	92	90.742	1.258
32	87	3	91	90.386	0.614
33	89	2	97	89.742	7.258
34	89	3	98	90.742	7.258
35	93	2	97	90.454	6.546
36	96	3	97	91.988	5.012
37	71	3	89	87.538	1.462
38	69	3	87	87.182	-0.182
39	73	2	88	86.894	1.106
40	76	0	76	85.428	-9.428
41	79	2	83	87.962	-4.962

Table 12 Predicted vs Actual 2nd marking period averages

Assumptions

Having completed the regression model, the assumptions of linear model must be tested. Table 13 shows the Durbin-Watson value of 1.16 is greater than one meaning the residuals are independent.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.689[a]	.475	.451	4.816	1.160

a. Predictors: (Constant), Test 1, Hours doing hw

b. Dependent Variable: 2nd mp average

Table 13 model summary

A review of multi-collinearity reveals a problem. Multicollinearity occurs when predictors are correlated with other predictors in the model making it difficult to have confidence in the results of the study. Since the values in Table 14 in the Tolerance column of 0.946 for hours doing homework and 0.946 for test one are greater than 0.9 the variables are too closely correlated for a confident model. Collinearity assumption is not fulfilled. Corrective action must be taken and another model created.

Coefficients[a]

Model		Unstandardized Coef		Standardized Coef	t	Sig.	95.0% Confidence Interval for B		Tolerance
		B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	68.941	4.771		14.449	.000	59.331	78.552	
	Hours doing hw	3.029	.624	.539	4.851	.000	1.771	4.287	.946
	Test 1	.178	.061	.322	2.901	.006	.054	.301	.946

a. Dependent Variable: 2nd mp average

Table 14 coefficients

Table 13 reflects good data because the variance proportions are high in one dimension and low in another dimension. Also both independent variables are not high in the same dimension.

Table 14 will assist in identification of outliers. Review of the Cooks distance value of 0.227 reveal no outliers since the value is less than one.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	76.59	98.67	88.67	4.479	48
Std. Predicted Value	-2.696	2.234	.000	1.000	48
Standard Error of Predicted Value	.701	2.696	1.127	.428	48
Adjusted Predicted Value	77.09	100.36	88.69	4.645	48
Residual	-11.993	9.247	.000	4.712	48
Std. Residual	-2.490	1.920	.000	.978	48
Stud. Residual	-2.595	2.038	-.002	1.018	48
Deleted Residual	-13.023	10.412	-.021	5.115	48
Stud. Deleted Residual	-2.783	2.115	-.006	1.039	48
Mahal. Distance	.015	13.752	1.958	2.692	48
Cook's Distance	.000	.227	.030	.055	48
Centered Leverage Value	.000	.293	.042	.057	48

a. Dependent Variable: 2nd mp average

Table 14 residual statistics

Collinearity Diagnostics[a]

				Variance Proportions		
Model	Dimension	Eigenvalue	Condition Index	(Constant)	Hours doing hw	Test 1
1	1	2.803	1.000	.00	.03	.00
	2	.186	3.878	.02	.96	.02

	3	.011	16.323	.98	.01	.98

a. Dependent Variable: 2nd mp average

Table 13 collinearity diagnostics

Independence of errors was checked by creating a residuals vs fits plot as listed in figure 2. The pattern of data reflects homoscedasticity which indicates that the residuals have constant variance. The standardized predicted y values are plotted on the x axis and the standardized residuals are plotted on the y axis. The goal is to insure that standardized residuals remain constant as predicated values increase.

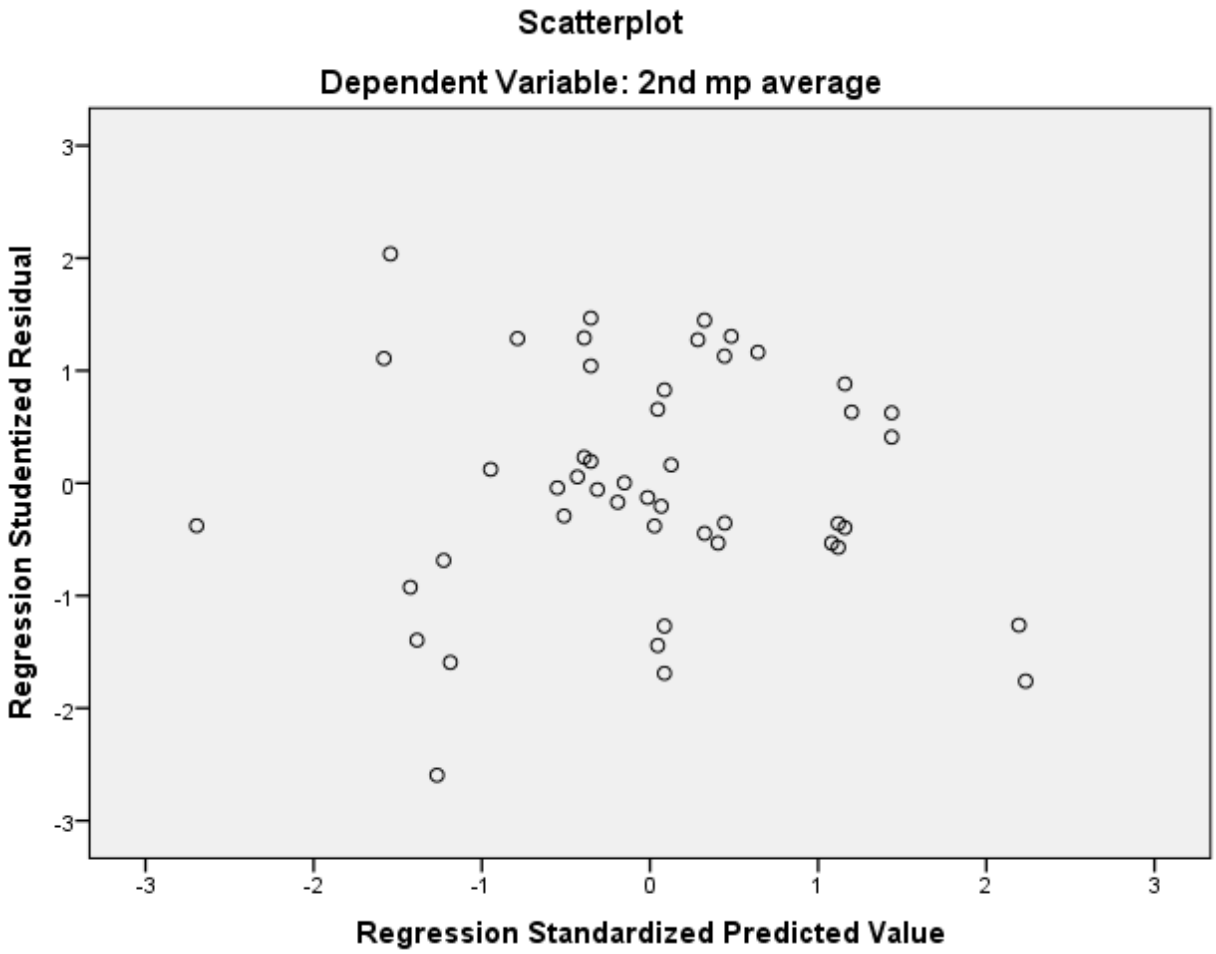

Figure 2 Residuals vs fits plot for model 2.

The normal P_P plot of regression sanitized residual shown in figure 3 demonstrates a linear relationship between the dependent and independent variables.

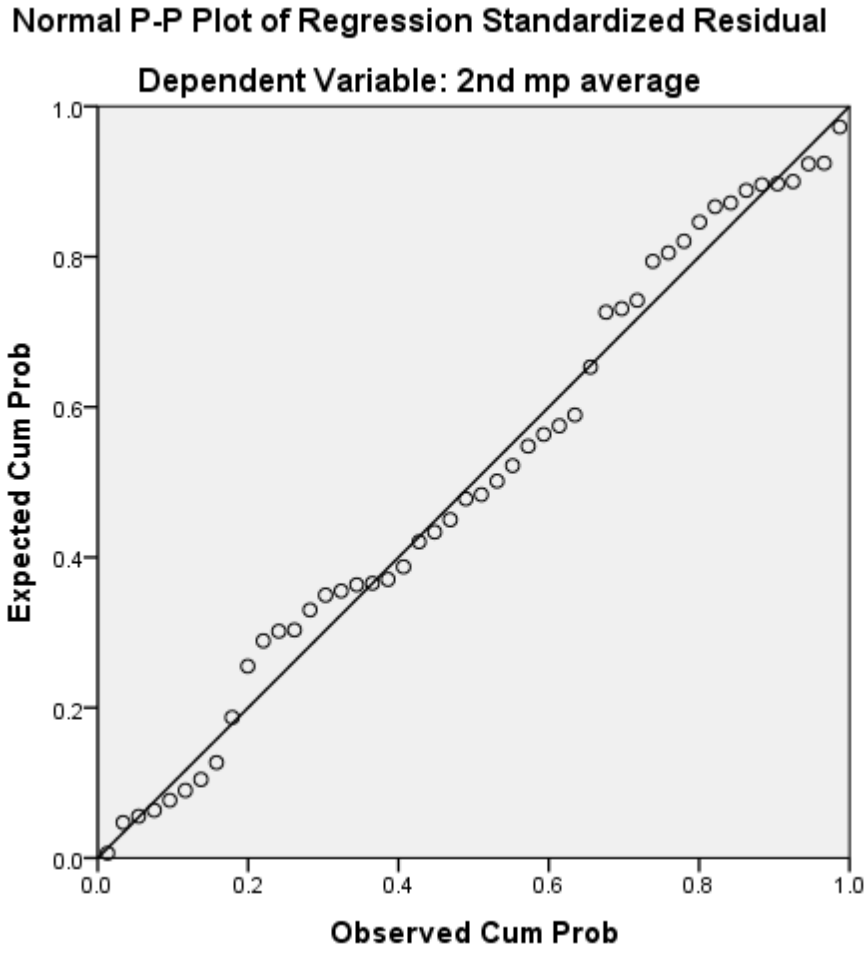

Figure 3 P-P plot of regression model 2.

Figure 4 below demonstrates the data is roughly normally distributed. Normality assumptions are fulfilled.

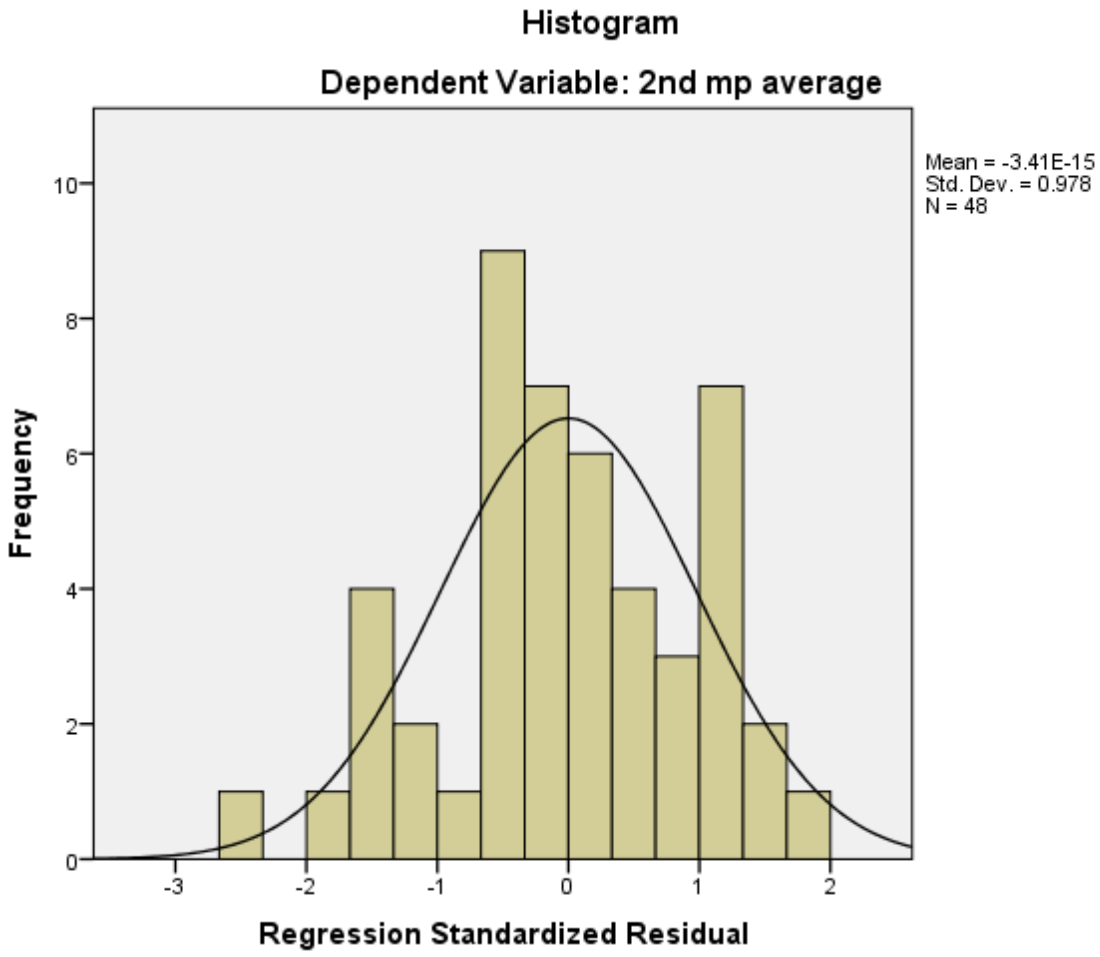

Figure 4 Histogram of standardized residuals

Conclusion

This issues of multicollinearity creates doubt on the strength of the model. Additional analysis comprising of but not limited to additional research and implementing different variables into the regression could possibly provide an effective model for determining factors that affect student's success in mathematics class. Upon completion of further analysis a more definitive conclusion could be developed.

References

Alcantara, A. (2014). Presentation of data. *Education.* Retrieved on March 31, 2016 from http://www.slideshare.net/andrilynalcantara/chapter-3-prsentation-of-data.

Executive Office of the President. (2010). Prepare and inspire: K-12 education in science, technology, engineering, and math (stem) for America's future. *President's Council of Advisors on Science and Technology*. Retrieved on April 27, 2016 from https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-stemed-report.pdf.

Appendix A

2nd MP avg	After School 1=yes	Math level 1=stand	Test 1	Hours Homework
91	1	0	55	1
91	0	0	57	2
85	0	0	70	1
88	1	0	70	3
88	1	1	70	3
87	0	1	72	2
78	0	0	75	0
82	0	1	78	2
93	0	1	79	2
75	1	0	43	0
80	1	0	80	0
86	0	0	80	1
90	1	0	80	2
93	0	1	81	5
88	0	1	84	1
96	0	0	84	2
86	1	1	37	2
88	0	1	85	2
88	0	0	85	2
92	0	1	85	1
94	0	1	85	1
87	0	0	86	1
91	1	0	88	3
92	0	1	88	3
96	1	1	88	2
97	1	1	90	3
	0	1	92	2
98	1	1	96	3
93	0	1	84	1
88	0	0	85	1