Multiple Regression on Student Performance Data Set

Statistical Methods I Project Report

Scott Williams & Lakesha Frost

1 Introduction

1.1 Data

The data set used to develop a multiple regression model to predict student performance was sourced from kaggle. Student marks, study time, and number of courses are the three variables present in the data set with a sample size of 100. Unfortunately, the author of the data set did not give many details about the units or scale of how the variables are measured. Number of courses is self explanatory and we believe it is reasonable to assume study time is measured in hours; otherwise, the maximum value of study time would imply the longest a student studied was 7.96 minutes which does not make sense. Student marks is the variable that we felt needed more explanation. On a traditional grading scale one would expect student marks to range from around 0 to 100. However, this data ranges from 5.61 to 55.3. We believe that the data was either scaled somehow or that we are working with performance on an assessment that was not graded on a traditional grading scale. We attempted to inspect this further since the author of the data set on kaggle stated that they downloaded the data set from the UCI Machine Learning Repository. While we did find data closely related to student performance on the UCI Machine Learning Repository, unfortunately, we could not find the exact data set. The UCI Machine Learning Repository data sets on student performance all were much larger with far more variables. The only way we could think to explain this is that the author on kaggle might have made their own data set out of a smaller section of a data set on the UCI Machine Learning Repository.

1.2 Goal of the Project

As students, we are often asked questions like how many classes are you taking this semester or how long did you study for the test. We wanted to explore the relationship between student performance, number of courses taken, and study time. Furthermore, we wanted to build a model that could predict a student's performance based on how many classes they are taking and how long they study with high predictive accuracy.

2 Preparation & Approach

2.1 Programming Environment & Preparation

All the code for this project was done in the SAS OnDemand for Academics Environment. The data set was downloaded from kaggle as a .csv file. Using proc import we loaded our data into SAS. The values for our data were stored as the Marks, number_courses, and time_study variables. We then split the data set 70/30 for model building and validation purposes. While splitting the data into the two different sets, random sampling was used to combat bias. The two data sets were stored as marks_build and marks_val.

2.2 Model Building - First Attempt

Since we are trying to build a model to predict student performance, Marks is our response variable while number_courses and time_study are our predictor variables. Before fitting a regression model to our data, we first wanted to get a feel for how our response variable was related to our predictor variables. We generated the scatter plot matrix for our data to visualize this relationship. We hoped to see the plots displaying linear patterns. As you can see in Figure 1, both predictor variables do appear to have an approximately linear relationship with our response variable. Observe that the plot of Marks vs time_study does have a slight curving pattern which we will explore further. The correlation matrix was also used to see how correlated our variables were to each other. From Figure 2 we see that both predictor variables are correlated to the response variable and have low correlation with each other. All of these results were what we hoped to see so we concluded our data was a good fit for a regression model.

Since we have multiple predictor variables, we needed to decide what variables to include in our model. We decided this using two different methods. Using adjusted R^2, we want to pick the model that has the highest adjusted R^2 value. As shown in Figure 3, that model is the model with all predictor variables included. As a way to confirm our results that the full model would be the best model, we also checked the results of using Forward Stepwise Regression. We found that this process also had the full model being the best model as shown in Figure 4.

With the model below being found as the best model, we fit it to our data.

$$Marks = \beta_0 + \beta_1 * number\_courses + \beta_2 * time\_study$$

We obtained the results shown in Figure 5. Our estimated regression function for the first attempt model was:

$$Marks = -8.52350 + 2.01017 * number\_courses + 5.51290 * time\_study$$

From the parameter estimates table in Figure 5, we can see that all estimated parameters are statistically significant based on their p-values. We also see that our model has a very high coefficient of multiple determination R^2. We thought these were all good signs that our model was performing well. However, upon inspecting the diagnostic plots, it is obvious that there are some problems.

On the residual plots for our model, we hoped to see a band around 0 with no obvious pattern and even width. This way we would be meeting the assumptions that the residuals were independent and had constant variance. However, on the Residuals vs Predicted Variable and Residuals vs time_study plots it is easy to see that these assumptions are not met. Both plots display a parabolic pattern. The number_courses plot fits the criteria of what we are hoping to see. Observe in Figure 7 that the QQ-Plot also has curvature which suggests our residuals are not normally distributed. We knew something wrong was going on with our model so we analyzed the partial regression plots to gain insight on how each variable was impacting the model. From the results displayed in Figure 8, we can see that again the number_courses plot looks good. It is displaying a linear pattern which indicates to us that we should add the linear term of number_courses to the model. However, the partial regression plot for time_study is again nonlinear. From the partial regression plots and the residuals plots, we decided we needed to transform the time_study variable.

2.3 Model Building - Second Attempt

The partial regression plot for time_study suggests we could have tried an exponential transform for the variable. However, based on the residual plot of the time_study variable in Figure 6 being shaped like a parabola, we transformed the variable to be time_study^2. From this point on, time_study^2 will be referred to as X_2. This is the variable name that we gave the squared value in SAS for coding simplicity. After transforming the time_study predictor variable, we wanted to inspect our scatter plot and correlation matrices to see if a regression model would be a good fit for our transformed data. We see in Figure 9 that the curving pattern in the Marks vs time_study has been corrected by the transform to display the linear pattern that we hope to see. In Figure 10 we see that the correlation matrix displays even higher levels of correlation between the response and predictor variables while keeping the correlation between the two predictor variables low. We felt these were good signs that the transformed data would work with a regression model.

We fit the following equation to our data:

$$Marks = \beta_0 + \beta_1 * number\_courses + \beta_2 * X_2$$

We obtained the results shown in Figure 11. Our estimated regression function for the transformed data was:

$$Marks = 0.39470 + 1.73480 * number\_courses + 0.66993 * X_2$$

We noticed that after performing the transform on time_study, the coefficient of multiple determination was very high. Since we want our R^2 value to be closer to 1, we were excited to see ours being 0.9997. After finding our estimated regression function, we wanted to check our diagnostic plots to see if the transformed variable corrected the problems in our model. Much like last time we first looked at our residual plots. In Figure 12, observe that all residual plots now look how we expect them to. They all display a band around 0 with no distinct pattern and constant variance. The QQ-plot in Figure 13 looks less curved than before. Although the plot isn't perfectly linear, we think based on the tighter linear pattern, it is fair to say that the residuals are approximately normally distributed. Finally, we looked to the partial regression plots in Figure 15 to see the effects of each variable on the model. Notice both plots are highly linear which means we should add both number_courses and X_2 to the model.

To gain further insight about the data used to build our model, we used SAS to identify outliers in both X and Y. In Figure 15, we can see that we have two outliers in Y and one outlier in X. The outliers in Y are cases 57 and 69. The outlying case in X is case 5. Since these cases are outliers, we wanted to see if they were also influential. We used Cook's Distance, DFFITS, and DFBETAS to determine if they were influential cases. In Figure 16, we see that Cook's Distance identified cases 57 and 69 to be influential. However, we have seen before that Cook's Distance can sometimes be too sensitive and identify cases as influential that are not considered to be by other metrics. Therefore, we decided to check both the DFFITS and DFBETAS values of these cases as well. To determine whether the cases were influential we used the following criteria where p is the number of parameters in our model and n is the sample size:

$$DFFITS\ Influential\ Value = 2 * \sqrt{\frac{p}{n}} = 2 * \sqrt{\frac{3}{70}} = 0.414$$

$$DFBETAS\ Influential\ Value = \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{70}} = 0.239$$

We compare the absolute value of the DFFITS and DFBETAS values for cases 57 and 69 and see if they are larger than the two above values. In Figure 17 observe that the values for cases 57 and 69 do meet the criteria to be considered influential.

Although it was quite obvious from both our scatter plot and correlation matrices, we wanted to do one final check for any multicollinearity problems. We checked the VIF values and obtained the results shown in Figure 18. Since our values are approximately equal to 1, we concluded that we do not have any multicollinearity problems.

Before moving on to validation, we wanted to make sure our model didn't need an interaction term between the two predictor variables. We performed the following hypothesis test with β_3 being the coefficient for the interaction term:

$$H_0 : \beta_3 = 0 \qquad H_1 : \beta_3 \neq 0$$

We used SAS to perform the test and obtained the results in Figure 19. Observe that our F Value = 0.58. The critical value that we compare this to is $f_{0.05,1,66} = 3.986$. We reject the null hypothesis if F ¿ than the critical value. Obviously, 0.58 is not greater than 3.986 so we fail to reject the null hypothesis and can drop the interaction term from the model.

2.4 Model Validation

After finalizing our model, we wanted to see how it would perform on a validation data set. There are multiple ways to validate a regression model. First, we used the same model form obtained during the building procedure on the validation data set. The second way was finding the MSPR value and comparing it to the MSE obtained during the model building stage. In Figure 20, we see that the model building estimated parameters (left) are fairly consistent with the validation estimated parameters (right). This is already a good sign that our model will generalize but we wanted to take it a step further and find the MSPR. In Figure 21, the MSPR value was found in SAS and is displayed in the mean column. Recall that the MSE obtained during the model building stage was 0.07337. We expect MSPR ¿ MSE. However, if the difference between MSPR and MSE of the model on the building data set is not noticeably different, it indicates good predictive ability. Here, we see that 0.0830572 and 0.07337 are similar in value. Therefore, the model has good predictive ability and does generalize well to new data.

3 Conclusion

3.1 Final Thoughts

We found it interesting that as both the number of courses and time studied increased so did student performance. Seeing the relationship between time studied and performance made sense to us but the relationship between performance and the number of courses taken surprised us. When we first found this data set, we thought that the number of courses would actually negatively impact a student's performance due to being more busy. The explanation for this goes beyond the scope of our data as there wasn't any extra information given about the students besides the performance, courses, and time studied values. However, one explanation may be that students who take more courses experience success in academics due to being more motivated students.

4 Figures and SAS Output

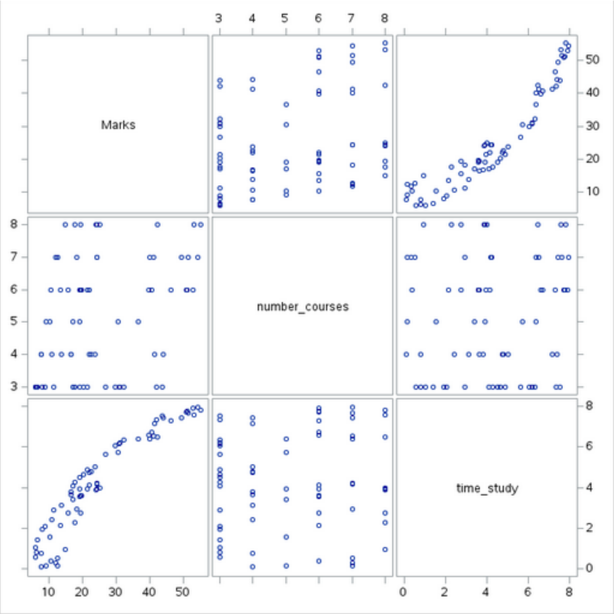
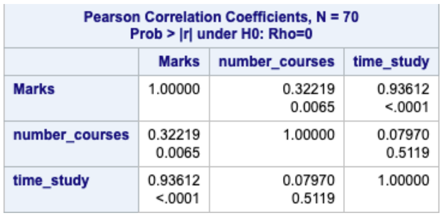


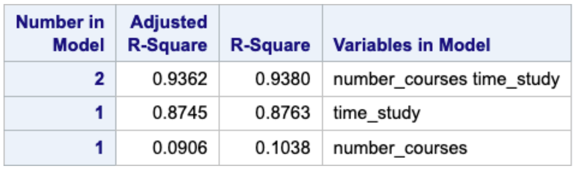
Figure 1: Scatter plot matrix for first attempt model building



Pearson Correlation Coefficients, N = 70
Prob > |r| under H0: Rho=0

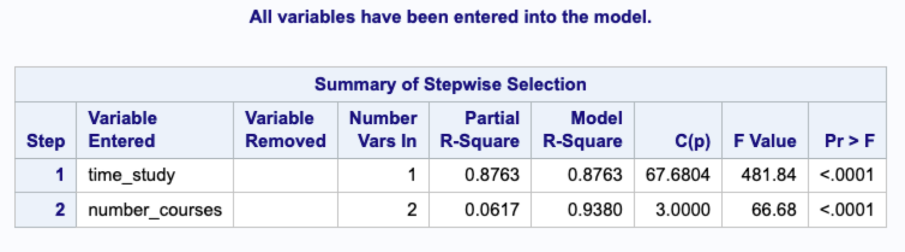
	Marks	number_courses	time_study
Marks	1.00000	0.32219 0.0065	0.93612 <.0001
number_courses	0.32219 0.0065	1.00000	0.07970 0.5119
time_study	0.93612 <.0001	0.07970 0.5119	1.00000

Figure 2: Correlation matrix for first attempt model building



Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.9362	0.9380	number_courses time_study
1	0.8745	0.8763	time_study
1	0.0906	0.1038	number_courses

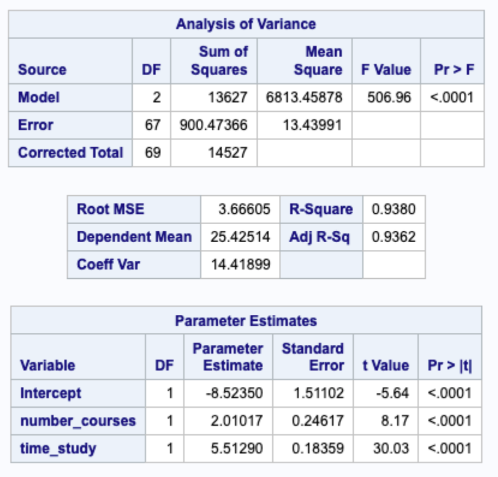
Figure 3: The resulting model after checking the adjusted R^2 criterion



All variables have been entered into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	time_study		1	0.8763	0.8763	67.6804	481.84	<.0001
2	number_courses		2	0.0617	0.9380	3.0000	66.68	<.0001

Figure 4: The resulting model after using Forward Stepwise Regression



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13627	6813.45878	506.96	<.0001
Error	67	900.47366	13.43991		
Corrected Total	69	14527			

Root MSE	3.66605	R-Square	0.9380
Dependent Mean	25.42514	Adj R-Sq	0.9362
Coeff Var	14.41899		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.52350	1.51102	-5.64	<.0001
number_courses	1	2.01017	0.24617	8.17	<.0001
time_study	1	5.51290	0.18359	30.03	<.0001

Figure 5: ANOVA table and estimated parameters for first attempt model

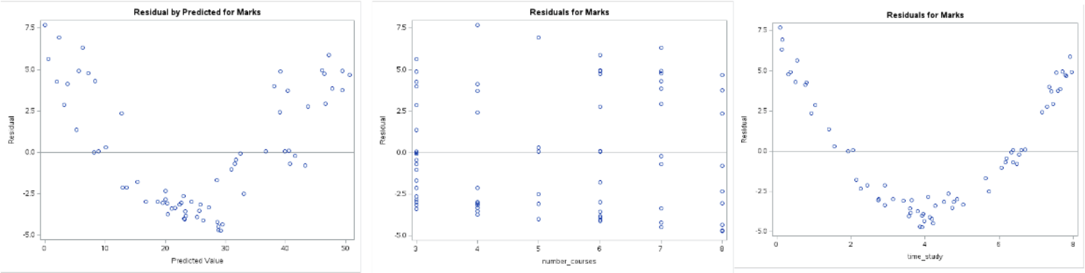


Figure 6: Diagnostic residual plots for first attempt model

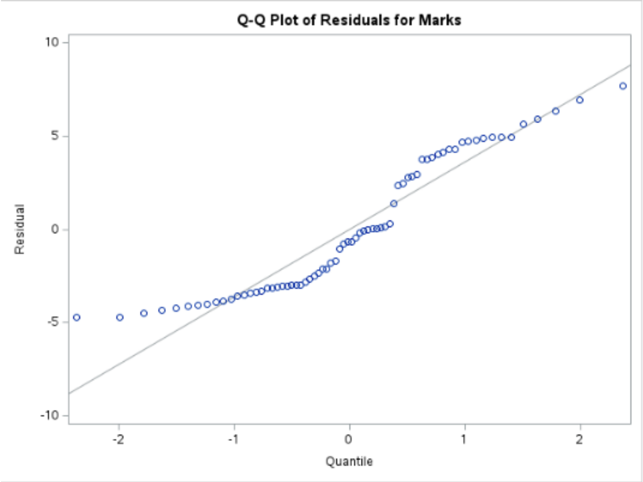


Figure 7: Diagnostic QQ-plot for first attempt model

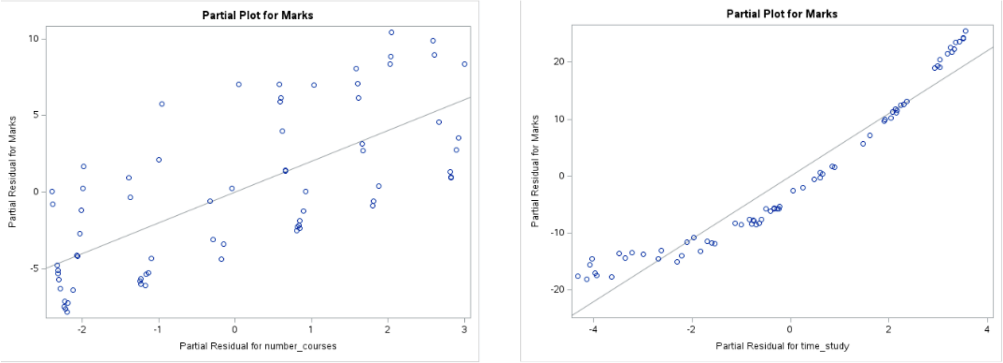


Figure 8: Diagnostic partial regression plots for first attempt model

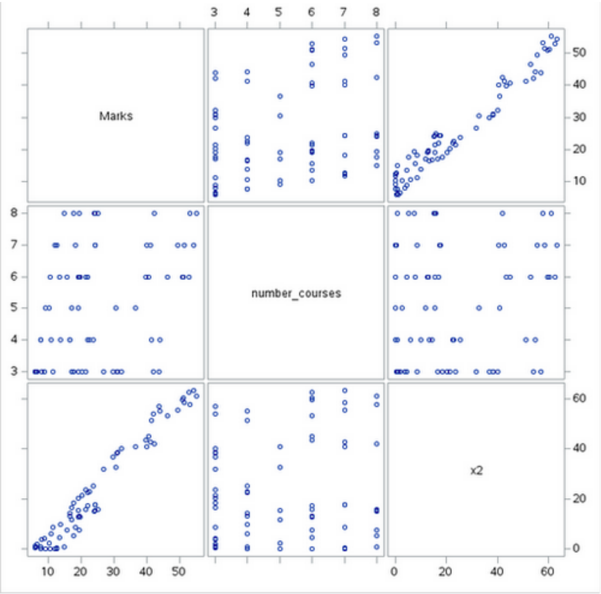
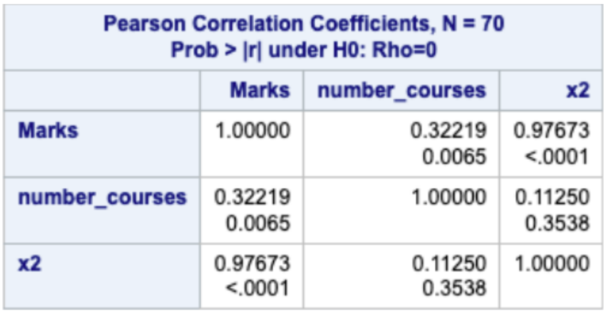


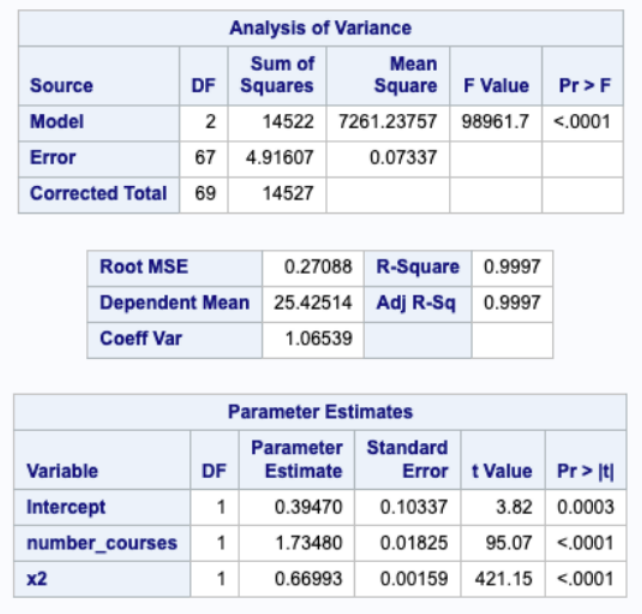
Figure 9: Scatter plot matrix of the transformed data



Pearson Correlation Coefficients, N = 70
Prob > |r| under H0: Rho=0

	Marks	number_courses	x2
Marks	1.00000	0.32219 0.0065	0.97673 <.0001
number_courses	0.32219 0.0065	1.00000	0.11250 0.3538
x2	0.97673 <.0001	0.11250 0.3538	1.00000

Figure 10: Correlation matrix of the transformed data



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14522	7261.23757	98961.7	<.0001
Error	67	4.91607	0.07337		
Corrected Total	69	14527			

Root MSE	0.27088	R-Square	0.9997
Dependent Mean	25.42514	Adj R-Sq	0.9997
Coeff Var	1.06539		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.39470	0.10337	3.82	0.0003
number_courses	1	1.73480	0.01825	95.07	<.0001
x2	1	0.66993	0.00159	421.15	<.0001

Figure 11: ANOVA table and estimated parameters for the transformed data

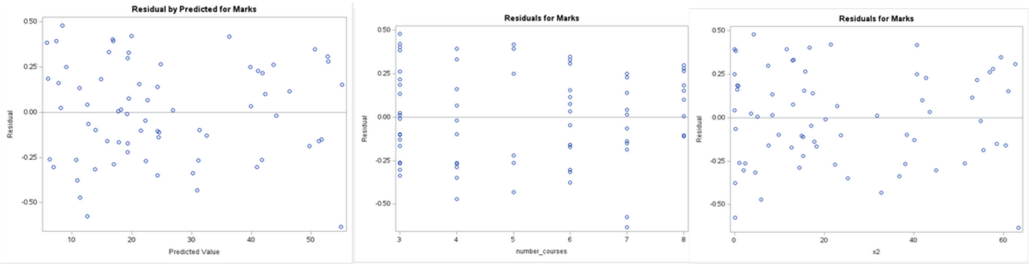


Figure 12: Diagnostic residual plots for the transformed data set

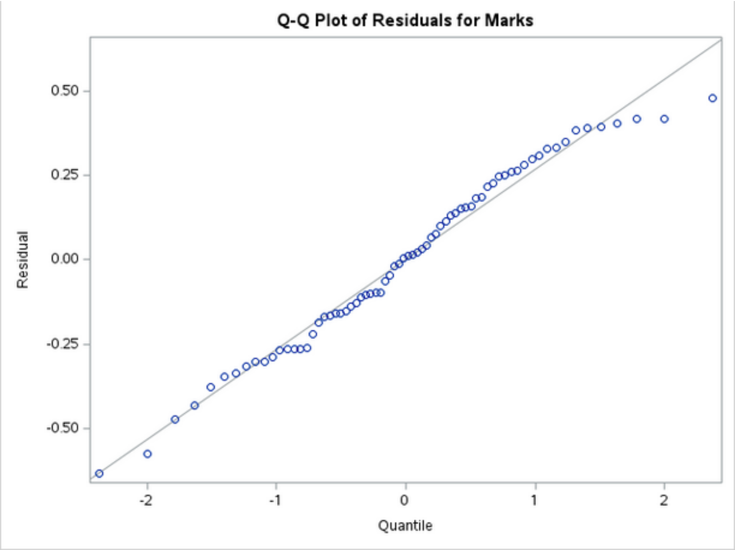


Figure 13: Diagnostic QQ-plot for the transformed data set

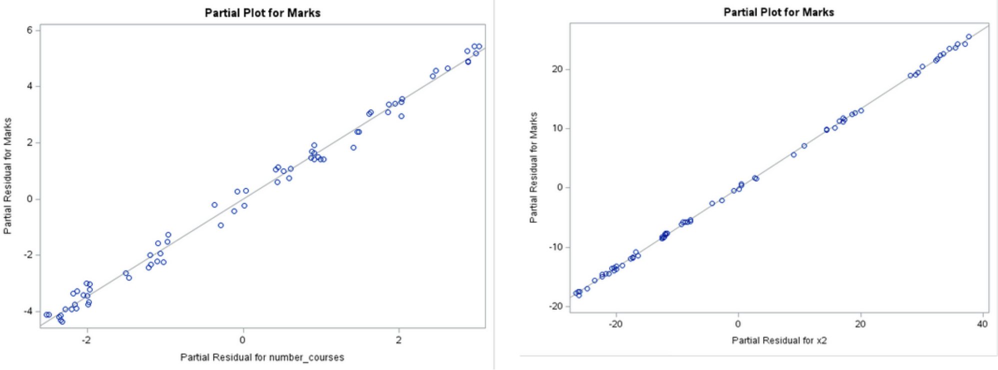


Figure 14: Diagnostic partial regression plots for the transformed data set

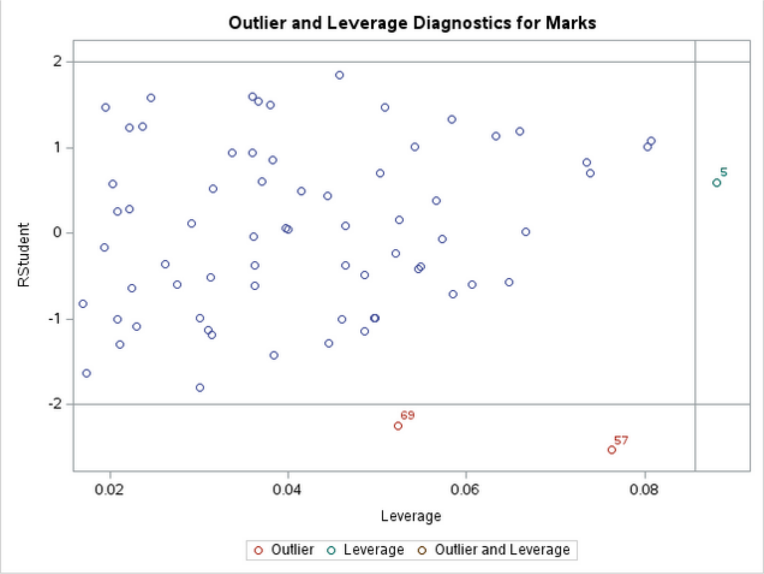


Figure 15: Outlier plot for transformed data set

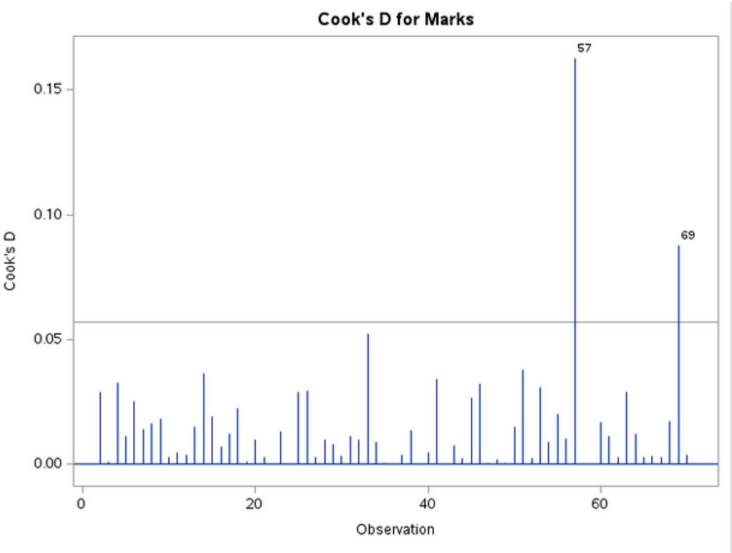
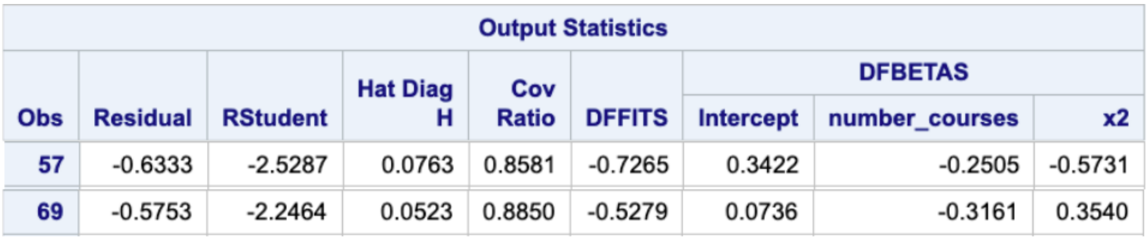
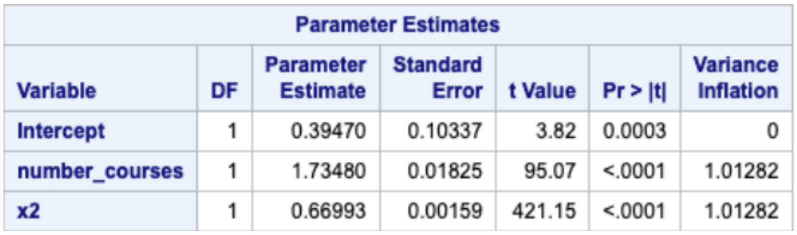


Figure 16: Cook's Distance plot for transformed data set


Output Statistics

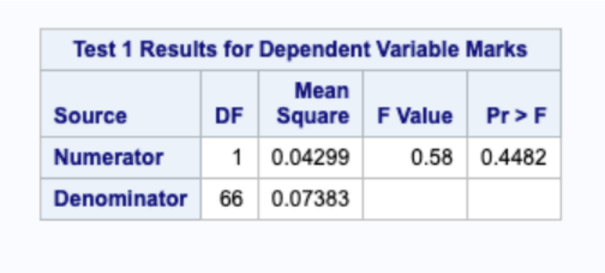
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
						Intercept	number_courses	x2
57	-0.6333	-2.5287	0.0763	0.8581	-0.7265	0.3422	-0.2505	-0.5731
69	-0.5753	-2.2464	0.0523	0.8850	-0.5279	0.0736	-0.3161	0.3540

Figure 17: DFFITS & DFBETAS values for the transformed data set's influential cases



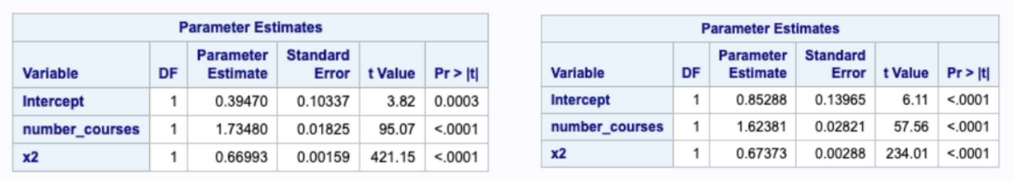
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.39470	0.10337	3.82	0.0003	0
number_courses	1	1.73480	0.01825	95.07	<.0001	1.01282
x2	1	0.66993	0.00159	421.15	<.0001	1.01282

Figure 18: VIF values for the transformed data set



Test 1 Results for Dependent Variable Marks				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.04299	0.58	0.4482
Denominator	66	0.07383		

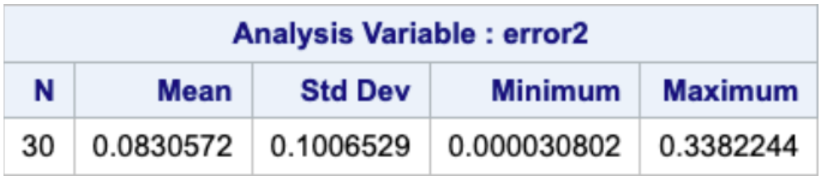
Figure 19: Hypothesis test results for the interaction term



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.39470	0.10337	3.82	0.0003
number_courses	1	1.73480	0.01825	95.07	<.0001
x2	1	0.66993	0.00159	421.15	<.0001

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.85288	0.13965	6.11	<.0001
number_courses	1	1.62381	0.02821	57.56	<.0001
x2	1	0.67373	0.00288	234.01	<.0001

Figure 20: Model building estimated parameters vs model validation estimated parameters



Analysis Variable : error2				
N	Mean	Std Dev	Minimum	Maximum
30	0.0830572	0.1006529	0.000030802	0.3382244

Figure 21: MSPR of the model