Wakjira Ashenafi

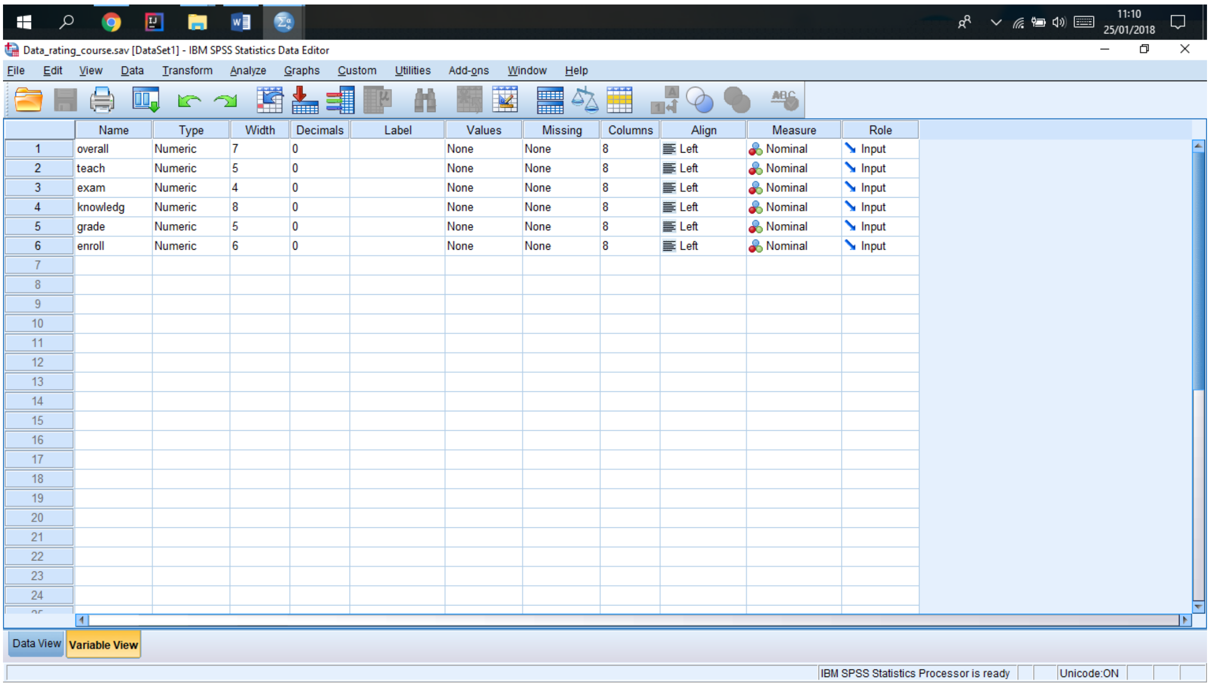
Data Analytics Software Project 1 Report

Multiple Regression Analysis With SPSS

Date 25 Jan. 18

Introduction

In this project we are given a dataset to predict an overall rating of a course using a multiple regression analysis on SPSS. The data set has 5 independent variables and one dependent variable with 50 records. Out of the five independent variables are required to use three of the independent variables named “teach” which is the teaching skill, “knowledge” which is the teacher’s knowledge and “grade” to predict the dependent variable “overall”.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	overall	Numeric	7	0		None	None	8	Left	Nominal	Input
2	teach	Numeric	5	0		None	None	8	Left	Nominal	Input
3	exam	Numeric	4	0		None	None	8	Left	Nominal	Input
4	knowledg	Numeric	8	0		None	None	8	Left	Nominal	Input
5	grade	Numeric	5	0		None	None	8	Left	Nominal	Input
6	enroll	Numeric	6	0		None	None	8	Left	Nominal	Input

Figure1. Variable’s tab overview of the data set on SPSS interface

In multiple regression, given two or more independent variables we want to determine their contribution to a single dependent variable. In this dataset taking the three-independent variable “teach”, “knowledge” and “grade”, we investigate their contribution to the independent variable “overall”.

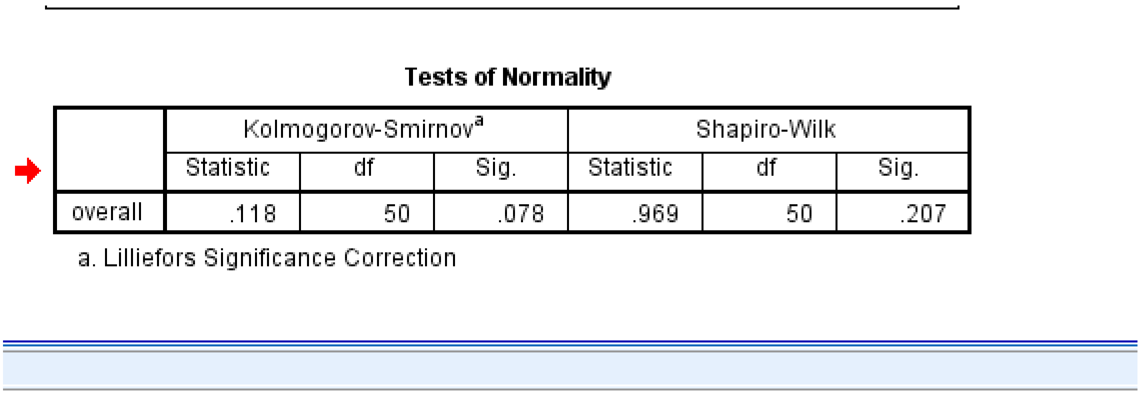
There are several assumptions while using multiple regression,

- Sample size – at least 20 records for each independent if the dependent variable is normally distributed otherwise we need to have more than 20 records for each independent variable
- Absence of outliers in all variables

- Linear relation between dependent and independent variables
- Absence of multicollinearity between independent variables

In this dataset we are given only 50 records, if we are considering the three independent variables “teach”, “knowledge” and “grade” contribution to the “overall” variable, we at least need 60 records to use multiple regression.

With this drawback let us continue checking the other assumptions. Let us check first the normality of the dependent variable “overall”. In SPSS this can be done **Analyze - > Descriptive Statistics -> Explore.** And on plots we check-off the descriptive histogram and normality plots with test. From the output window we are interested on test of normality.



Tests of Normality

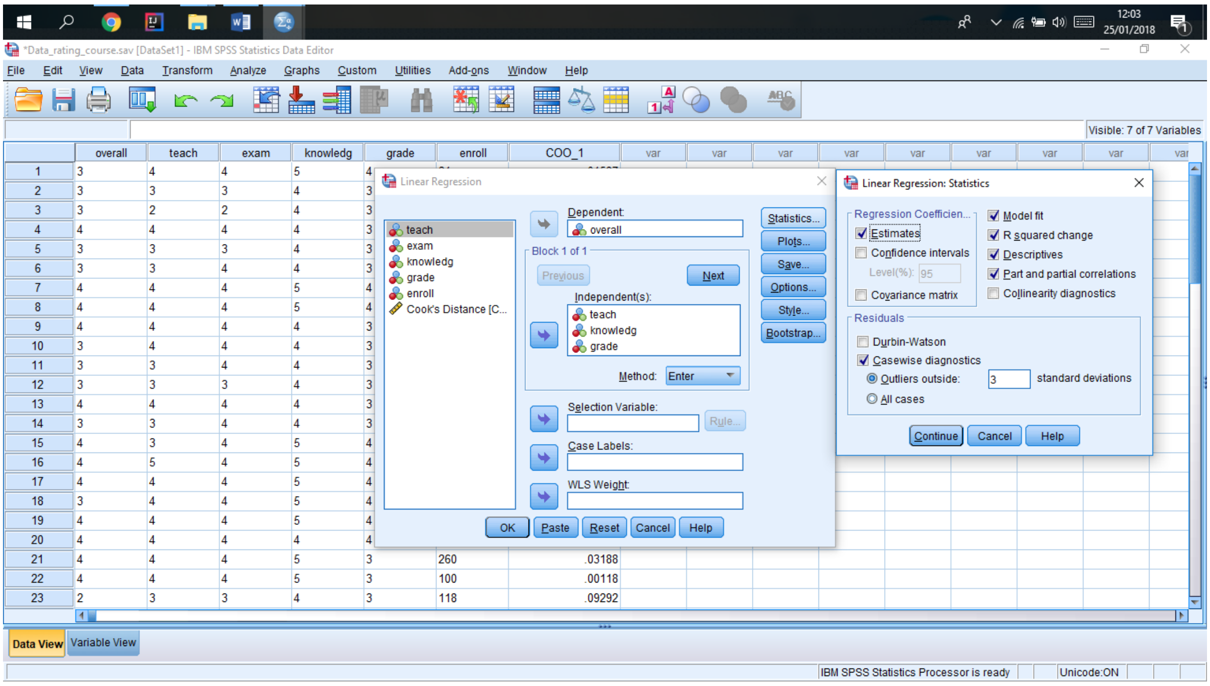
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
overall	.118	50	.078	.969	50	.207

a. Lilliefors Significance Correction

Table1. Tests of normality table

Here we are looking for a non-statistically significant result, so that we can assume a normally distributed variable. To check that we are looking the **Sig.** values in Shapiro- Wilk or Kolmogorov-Simirnov. If the value is greater than 0.005 then we have a non-statistically significant variable. In both cases Sig value (0.207 and 0.078) are greater than 0.005. So, we can conclude that the dependent variable “Overall” is normally distributed.

The other assumptions can be checked in the linear regression procedure. To do that in SPSS. Go to **Analyze - > Regression -> Linear.** Put the variables in their respective places, “teach”, “knowledge” and “grade” in independent box and “overall” in the dependent box. And then we set all the necessary setting on each field, statistics, plots, options etc as in the following picture.



*Data_rating_course.sav [DataSet1] - IBM SPSS Statistics Data Editor

Visible: 7 of 7 Variables

	overall	teach	exam	knowledg	grade	enroll	COO_1
1	3	4	4	5	4		
2	3	3	3	4	3		
3	3	2	2	4	3		
4	4	4	4	4	3		
5	3	3	3	4	3		
6	3	3	4	4	3		
7	4	4	4	5	4		
8	4	4	4	5	4		
9	4	4	4	4	3		
10	3	4	4	4	3		
11	3	3	4	4	3		
12	3	3	3	4	3		
13	4	4	4	4	3		
14	3	3	4	4	3		
15	4	3	4	5	4		
16	4	5	4	5	4		
17	4	4	4	5	4		
18	3	4	4	5	4		
19	4	4	4	5	4		
20	4	4	4	4	4		
21	4	4	4	5	3	260	.03188
22	4	4	4	5	3	100	.00118
23	2	3	3	4	3	118	.09292

Linear Regression

Dependent: overall

Block 1 of 1

Independent(s): teach, knowledg, grade

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Linear Regression: Statistics

Regression Coefficien...
- ☑ Estimates
- □ Confidence intervals Level(%): 95
- □ Covariance matrix
- ☑ Model fit
- ☑ R squared change
- ☑ Descriptives
- ☑ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- ☑ Casewise diagnostics
- ● Outliers outside: 3 standard deviations
- ○ All cases

Figure 2. How to set the interface in SPSS

After setting all the required fields we can now go to the output window and interpret our outs. From the descriptive Statistics we can observe the mean and standard deviation form 50 records. Since the data is collected as a questioner from 1 to 5, the data has no issue of normalization since all is on the same scale.

Descriptive Statistics

	Mean	Std. Deviation	N
overall	3.55	.614	50
teach	3.66	.532	50
knowledg	4.18	.408	50
grade	3.49	.351	50

Table 2. Descriptive Statistics table

Next let us check the rest of the assumptions, from the correlation table we can check the multicollinearity between the independent-independent variables. If the value of the Pearson correlation factor is greater than 0.7 then we conclude that the variables are multicollinear. If they are multicollinear then it is not ideal for further analysis. Since our assumption is that each variable is independent, the existence of the multicollinearity tells that the variables are giving redundant information. According to our correlation table below the Pearson correlation for all variables are less than 0.7. So, we can conclude that there is no multicollinearity between the independent variables.

Correlations

		overall	teach	knowledg	grade
Pearson Correlation	overall	1.000	.804	.682	.301
	teach	.804	1.000	**.526**	**.469**
	knowledg	.682	**.526**	1.000	**.224**
	grade	.301	**.469**	**.224**	1.000
Sig. (1-tailed)	overall	.	.000	.000	.017
	teach	.000	.	.000	.000
	knowledg	.000	.000	.	.059
	grade	.017	.000	.059	.
N	overall	50	50	50	50
	teach	50	50	50	50
	knowledg	50	50	50	50
	grade	50	50	50	50

Table 3. Correlations table

The next assumption to check is the linear relationship between the dependent and independent variable and there should be a linear relationship to go to further analysis. If the dependent-independent correlation is greater than 0.3 then we can conclude there exists a linear relation with them. In our case we can see that on the same table correlations. As we can see from the table the correlation is greater than 0.3.

The linearity of the dependent-independent variable can further have observed from the p-p plot below. As we can observe form the plot below the points lie on the line with some deviation. That tells that the independent-dependent variables are linearly correlated.

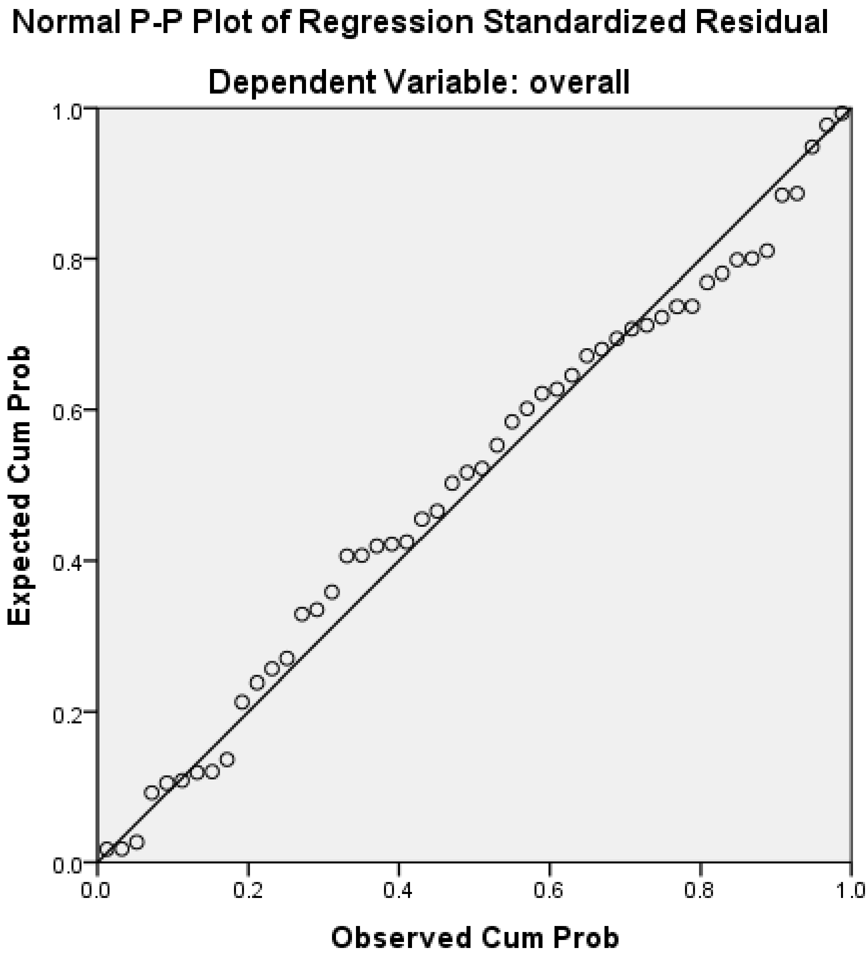


Figure 3. P-P plot of regression of expected com prob vs observed cum pro

Other points to check for example from the residual statistics table below is that standard residual and it must be between -3 to 3 of min and max value. And the Cook's distance shouldn't be greater than 1. And both values meet the conditions as shown below on the table.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.40	4.58	3.55	.529	50
Std. Predicted Value	-2.178	1.940	.000	1.000	50
Standard Error of Predicted Value	.048	.156	.087	.027	50
Adjusted Predicted Value	2.35	4.64	3.55	.535	50
Residual	-.675	.789	.000	.310	50
Std. Residual	-2.110	2.467	.000	.969	50

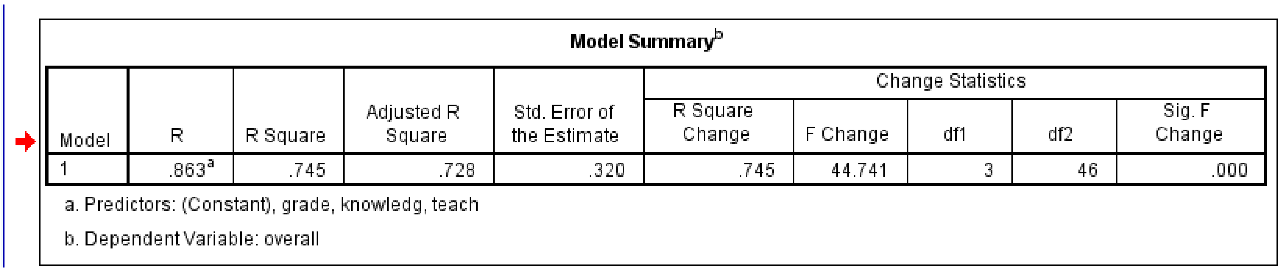
Stud. Residual	-2.191	2.719	.007	1.012	50
Deleted Residual	-.727	.958	.005	.338	50
Stud. Deleted Residual	-2.289	2.935	.007	1.040	50
Mahal. Distance	.116	10.717	2.940	2.519	50
Cook's Distance	.000	.396	.024	.058	50
Centered Leverage Value	.002	.219	.060	.051	50

a. Dependent Variable: overall

Table 4. Residual statistics table

Interpretation of the Model

From the model **summery table**, we can see the R square value, also called the coefficient of determination. It is the amount of variance expressed as percentage that is explained in the dependent variable by the independent variable. We must be careful when using the R square value especially when we have a small recodes since it over estimates the value. In our case we have only 50 records, so we must use the adjusted R square value for our interpretation of R square value. What the **adjusted R square** explains is that our model explains 72.8 % of the variance the dependent variable and it is statistically significant finding as we can see from the **Sig. F change** value which is 0.000. To be statistically significant the value of the Sig. F change must be less than 0.005



Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.863[a]	.745	.728	.320	.745	44.741	3	46	.000

a. Predictors: (Constant), grade, knowledg, teach

b. Dependent Variable: overall

Table 5. Model summary table

The ANOVA table also tales use the statistical significance of the hypothesis since Sig value is less than 0.05.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.737	3	4.579	44.741	.000[b]
	Residual	4.708	46	.102		
	Total	18.445	49			

a. Dependent Variable: overall

b. Predictors: (Constant), grade, knowledg, teach

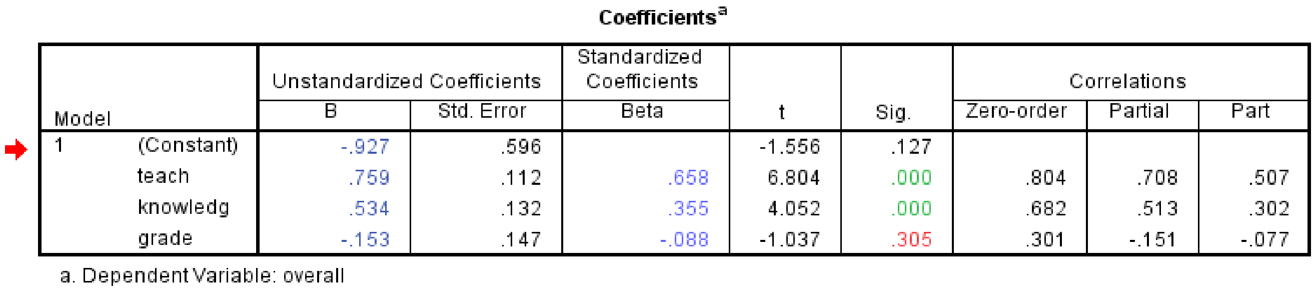
Table 6. The ANOVA table

From the coefficients table we can see the beta(B) value that is used in linear regression model. The unstandardized coefficients of B are the one used in the evaluation of the regression equation. In our case the linear equation will be

*Overall = -0.927 + 0.759*teach + 0.534*knowledge – 0.153*grade*

What it means is one unit change of teaching skill has 0.725 incremental change in overall variable, similarly one unit change on teacher's knowledge has 0.534 incremental change on overall variable, where as one unit change on grade has 0.153 decremental change on the variable overall and the constant -0.927 is the y-intercept of the equation.

But for comparison purpose we use the standardized coefficients of beta. As can been seen from the standardized coefficients of beta the **teach** (0.658) variable has the highest contribution than **knowledge** (0.355) and **grade** (-0.088) variables.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-.927	.596		-1.556	.127			
	teach	.759	.112	.658	6.804	.000	.804	.708	.507
	knowledg	.534	.132	.355	4.052	.000	.682	.513	.302
	grade	-.153	.147	-.088	-1.037	.305	.301	-.151	-.077

a. Dependent Variable: overall

Table 8. Coefficients table

We can show the contribution of each variables with a bar graph if needed like below,

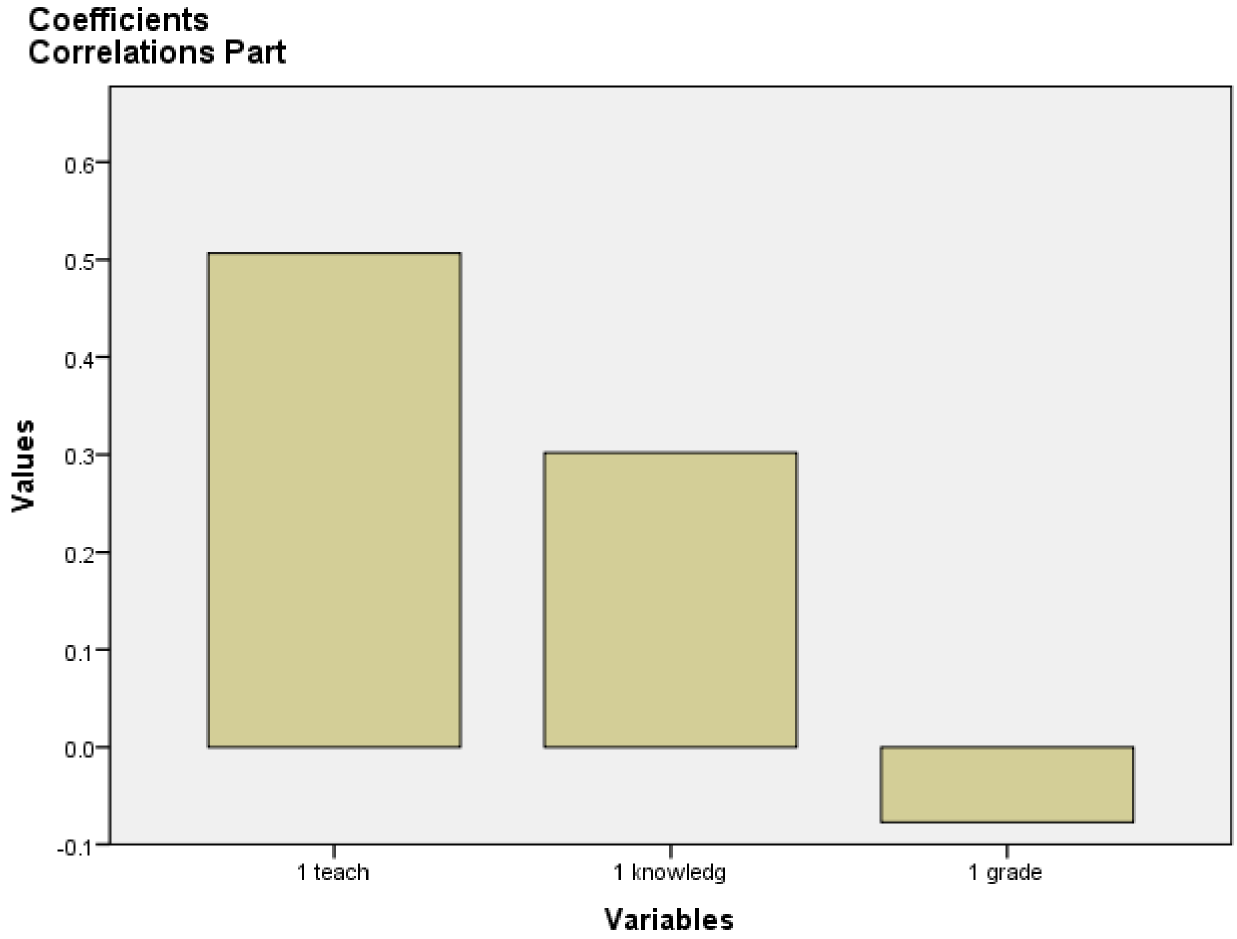


Figure 4. The unique contribution of each variables based on Part correlation.

We can also see the statistical significance of each variables from the same table. To check the significance, we have the p-values of 0.000 for **teach** and **knowledge** and 0.305 for **grade**. From those result we can conclude that teach and knowledge are statistically significant whereas **grade** is not since the value is greater than 0.05. As a result, we can say that grade is not a good predictor variable in this model.

On the same table, the part correlation tells the unique contribution of the variable, and we can note that the highest value is 0.507, which corresponds to teach variable.

Next, we can check the scatter plot of the dependent variable “overall” against the regression adjusted predicted value. The adjusted predicted value for case i is the predicted value that would be calculated for the case if the regression coefficients were estimated using all the other cases that were used in the current regression equation but case i omitted.

Formally,

$Adjusted\ predicted\ value_i = Y_i - DRESID_i$

Where $Adjusted\ predicted\ value_i$ is the adjusted predicted value for ith term, Y_i is observed value of dependent variable Y and $DRESID_i$ is deleted residual for Y in the ith term.

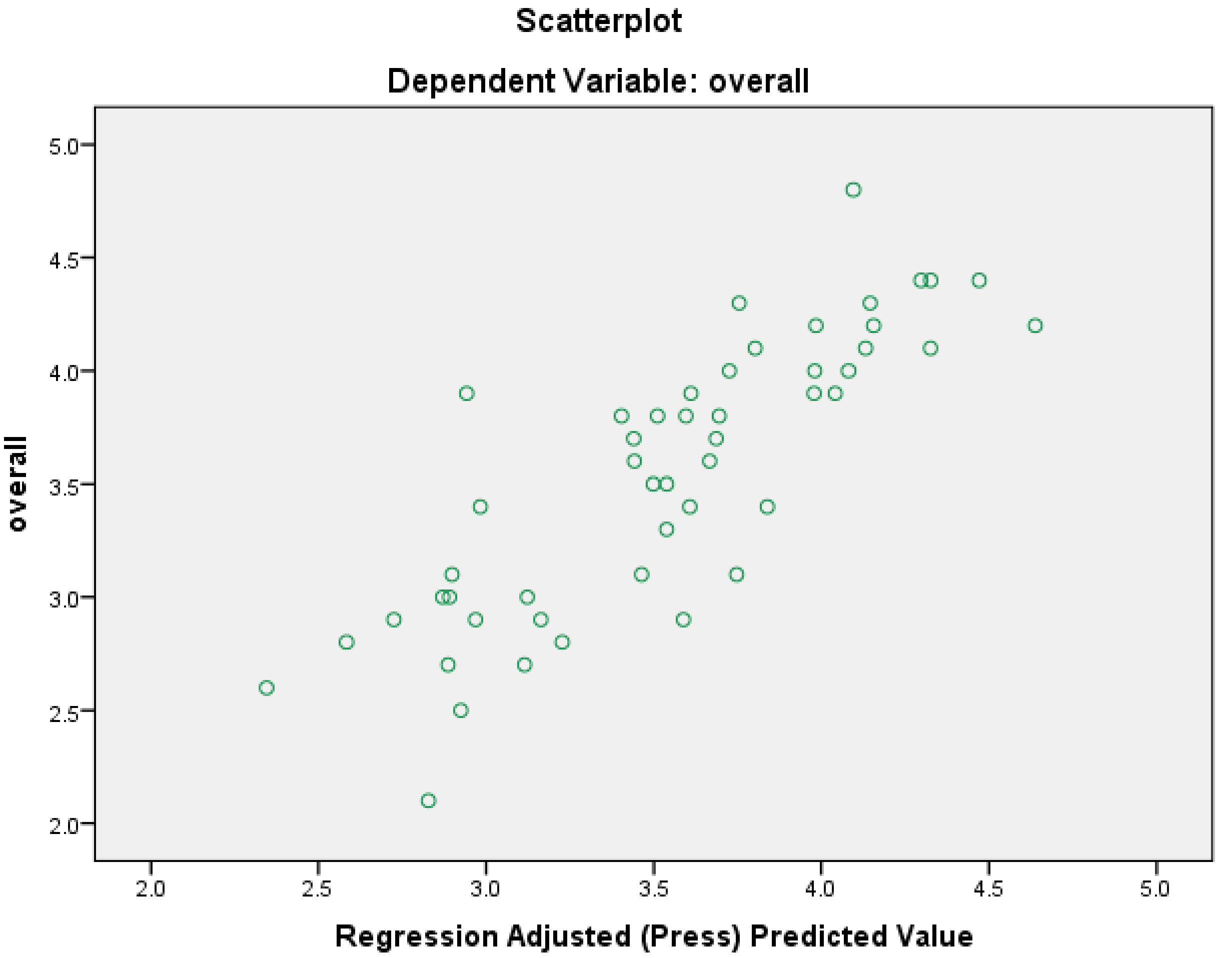


Figure 5. The dependent variable(overall) vs Regression Adjusted Predicted Value.

Conclusion and limitation

While we use multiple regression model on the desired data set, there are some assumptions that we take in to consideration as mentioned in the introduction part. One of the assumption was the data sample size. We need to have a reasonable amount of data to run machine learning algorithms. While using multiple regression at least we need to have a minimum of 20 records for each predictor variables. In this case, three predictor variables are used and minimum of 60 record is required to have unbiased result. Only 50 records are used in this case and this may result some biased result on the analysis. All results and interpretation are given with this limitation on the data.