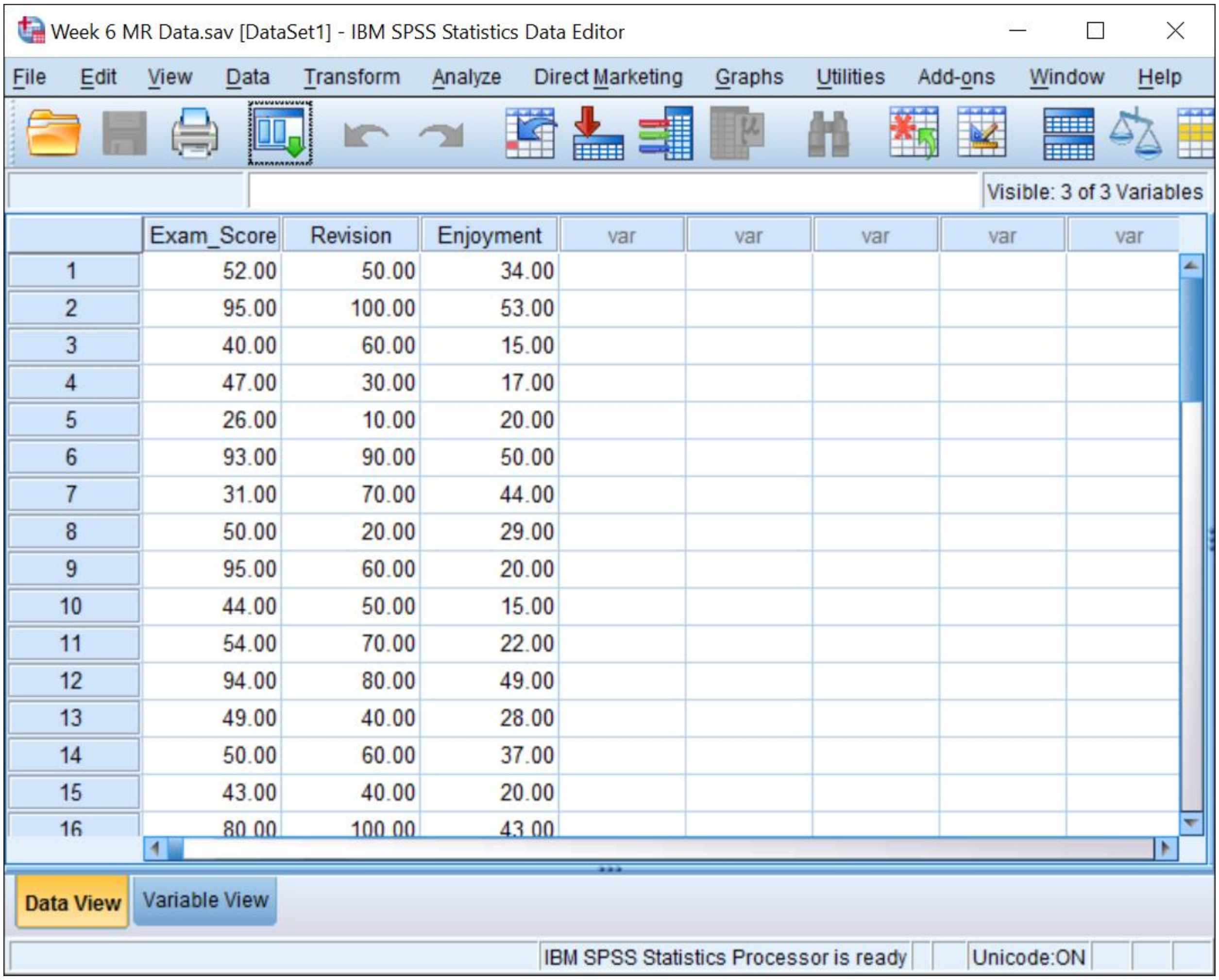
Assumptions of Multiple Regression

This tutorial should be looked at in conjunction with the previous tutorial on Multiple Regression. Please access that tutorial now, if you haven't already.

When running a Multiple Regression, there are several assumptions that you need to check your data meet, in order for your analysis to be reliable and valid. This tutorial will talk you though these assumptions and how they can be tested using SPSS.

This tutorial will use the same example seen in the Multiple Regression tutorial. It aims to investigate how revision intensity and subject enjoyment (the IVs/predictor variables) may predict students' exam score (the DV/outcome variable).

The analysis for this tutorial is all done using SPSS file 'Week 6 MR Data.sav':


Week 6 MR Data.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 3 of 3 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

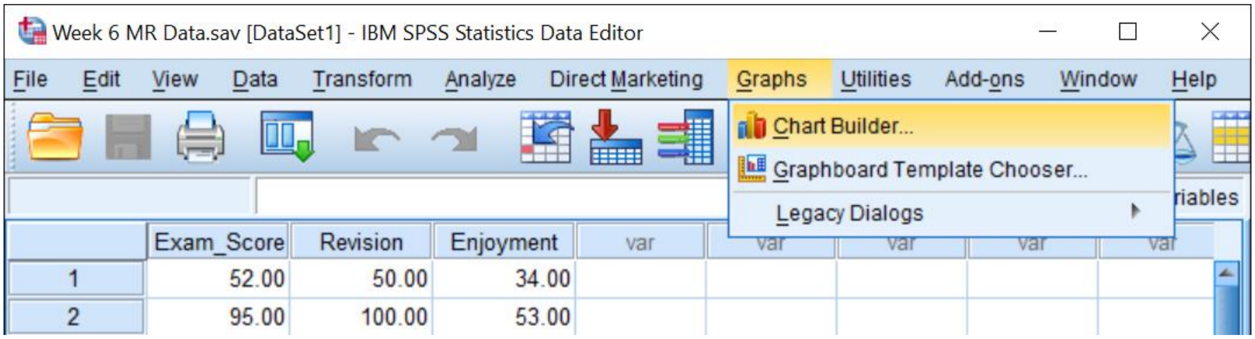
	Exam_Score	Revision	Enjoyment	var	var	var	var	var
1	52.00	50.00	34.00					
2	95.00	100.00	53.00					
3	40.00	60.00	15.00					
4	47.00	30.00	17.00					
5	26.00	10.00	20.00					
6	93.00	90.00	50.00					
7	31.00	70.00	44.00					
8	50.00	20.00	29.00					
9	95.00	60.00	20.00					
10	44.00	50.00	15.00					
11	54.00	70.00	22.00					
12	94.00	80.00	49.00					
13	49.00	40.00	28.00					
14	50.00	60.00	37.00					
15	43.00	40.00	20.00					
16	80.00	100.00	43.00					

The Assumptions

Assumption #1: The relationship between the IVs and the DV is linear.

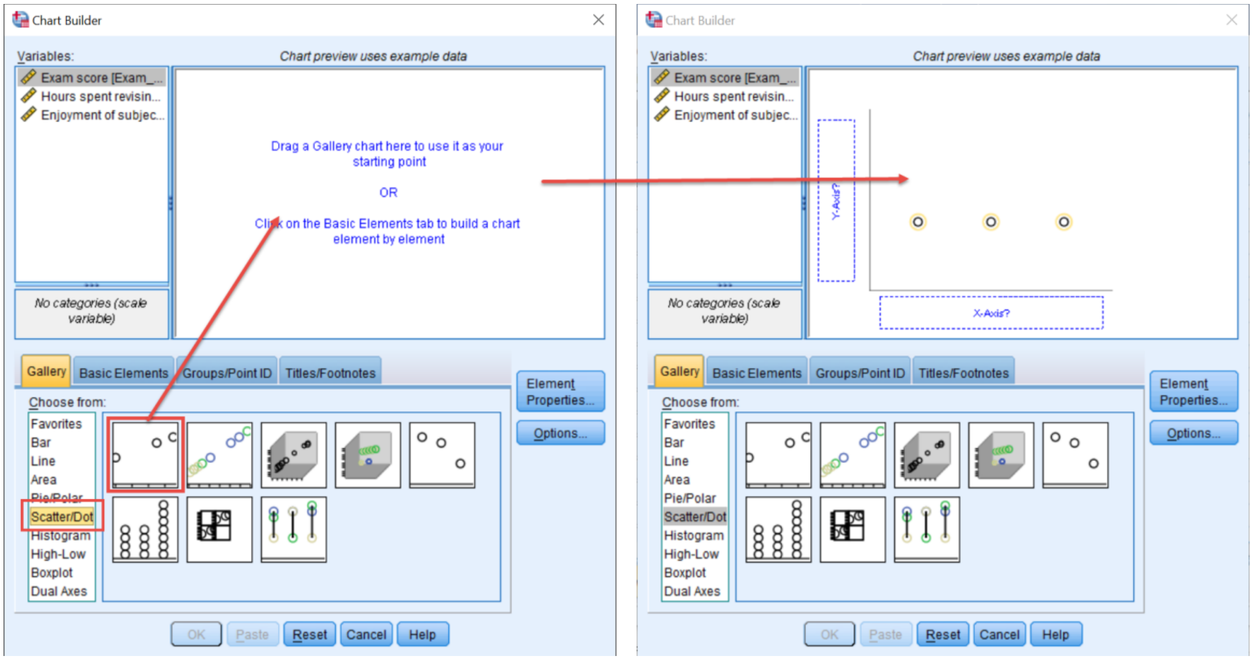
The first assumption of Multiple Regression is that the relationship between the IVs and the DV can be characterised by a straight line. A simple way to check this is by producing scatterplots of the relationship between each of our IVs and our DV.

To produce a scatterplot, **CLICK** on the **Graphs** menu option and **SELECT Chart Builder**

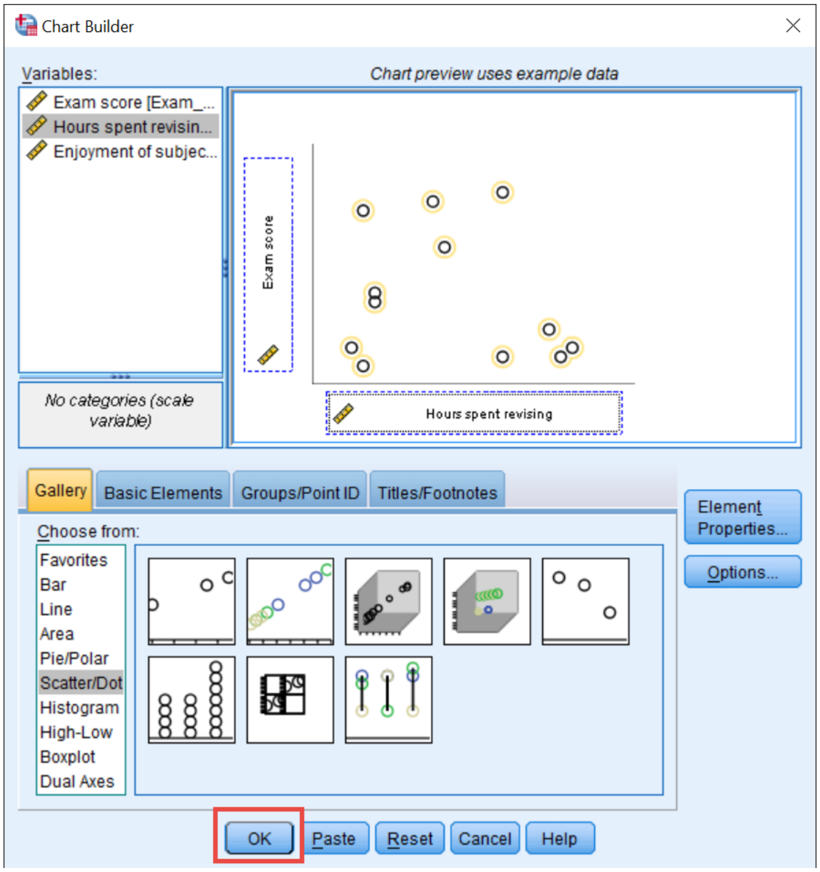


	Exam_Score	Revision	Enjoyment	var	var	var	var	var
1	52.00	50.00	34.00					
2	95.00	100.00	53.00					

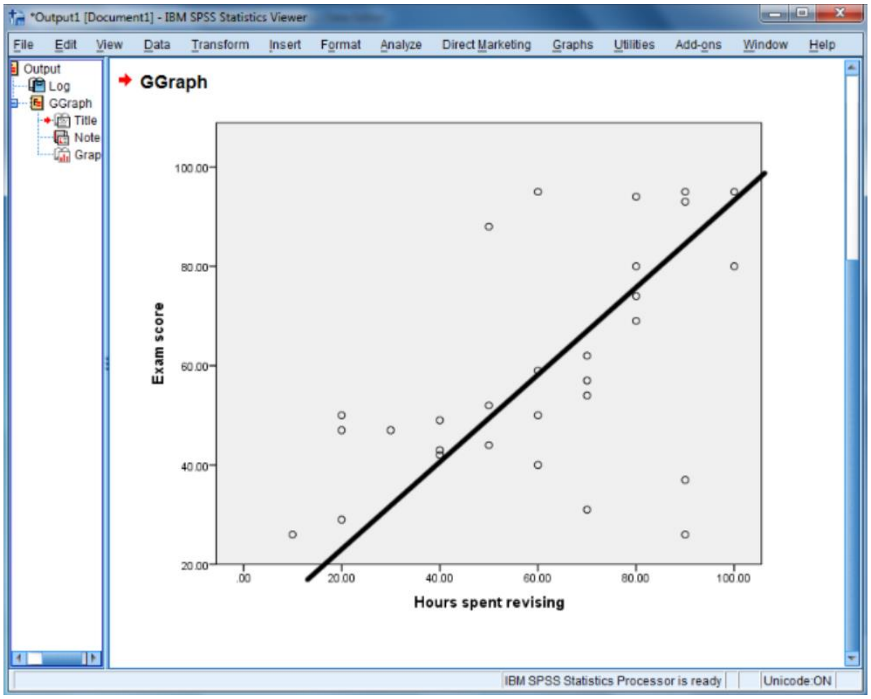
To produce a scatterplot, **SELECT** the **Scatter/Dot** option from the Gallery options in the bottom half of the dialog box. Then drag and drop the **Simple Scatterplot icon** into the **Chart Preview Window**.



Next, we need to tell SPSS what to draw. To do this, drag and drop the DV (**Exam Score**) onto the graph's Y-Axis and one of the IVs (in this case, **Hours spent revising**) onto the graph's X-Axis.



Now we have told SPSS what graph we want it to produce, <u>**CLICK**</u> on **OK** to draw the scatterplot



The graph can now be viewed in the Output Viewer.

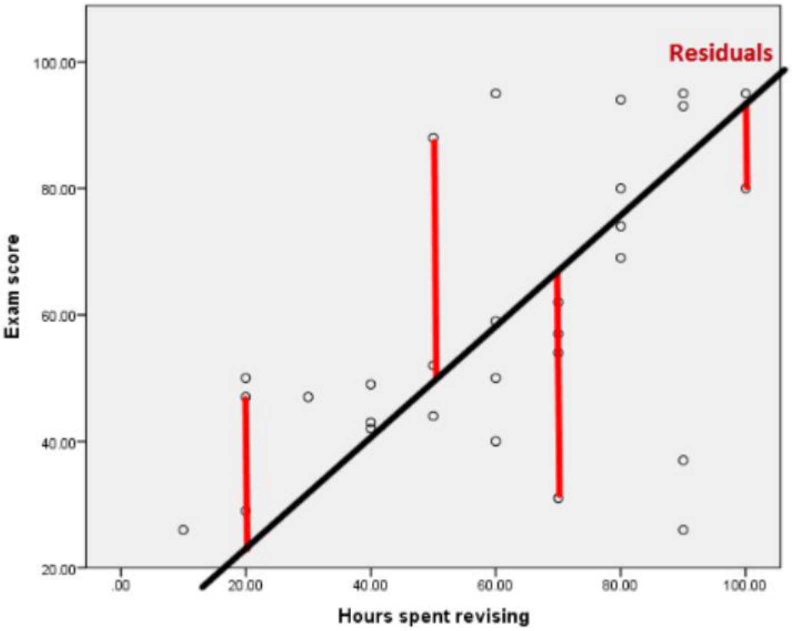
Looking at the scatterplot produced by SPSS, we can see that the relationship between the IV and the DV could be modelled by a straight line... suggesting that the relationship between these variables *is* linear.

But to fully test the assumption of linearity, you would need to do this for each of the IVs and the DV.

We have now looked at how to assess the first assumption of multiple regression. To understand several of the other assumptions, you first need to understand what is meant by the term ‘residuals’.

To explain this - look at the black line drawn on the graph. This represents a linear model of our data. We can see how well this line models our data by looking at how closely the different data points fall to the line. The closer they are, the more accurate the model.

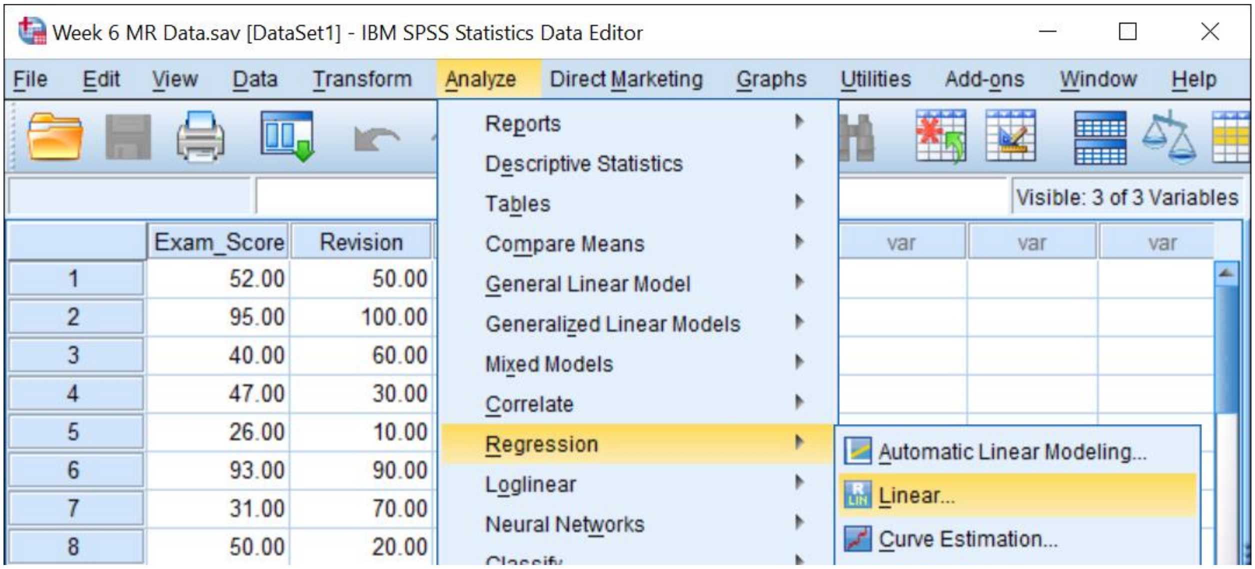
The vertical distance between the model and our data points represent the error in our model.



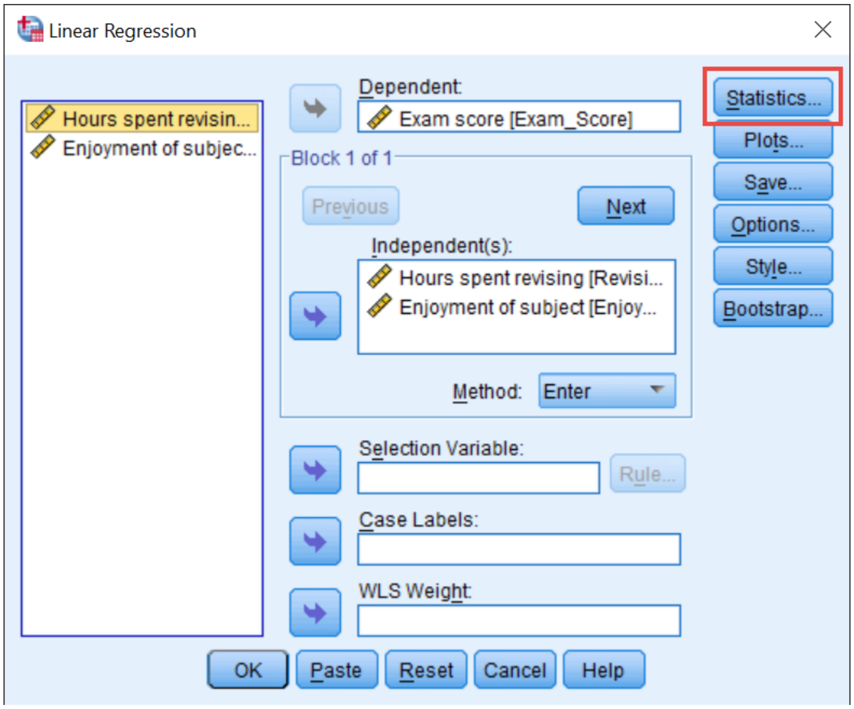
These distances are known as ***residuals***. Each data point has an associated residual, and these play an important role in the assumptions of multiple regression.

To test the next assumptions of multiple regression, we need to re-run our regression in SPSS.

To do this, **<u>CLICK</u>** on the **Analyze** file menu, **<u>SELECT</u> Regression** and then **Linear**.



	Exam_Score	Revision
1	52.00	50.00
2	95.00	100.00
3	40.00	60.00
4	47.00	30.00
5	26.00	10.00
6	93.00	90.00
7	31.00	70.00
8	50.00	20.00



Hours spent revisin...
Enjoyment of subjec...

Dependent: Exam score [Exam_Score]

Block 1 of 1

Independent(s):
Hours spent revising [Revisi...
Enjoyment of subject [Enjoy...

Method: Enter

Selection Variable:

Case Labels:

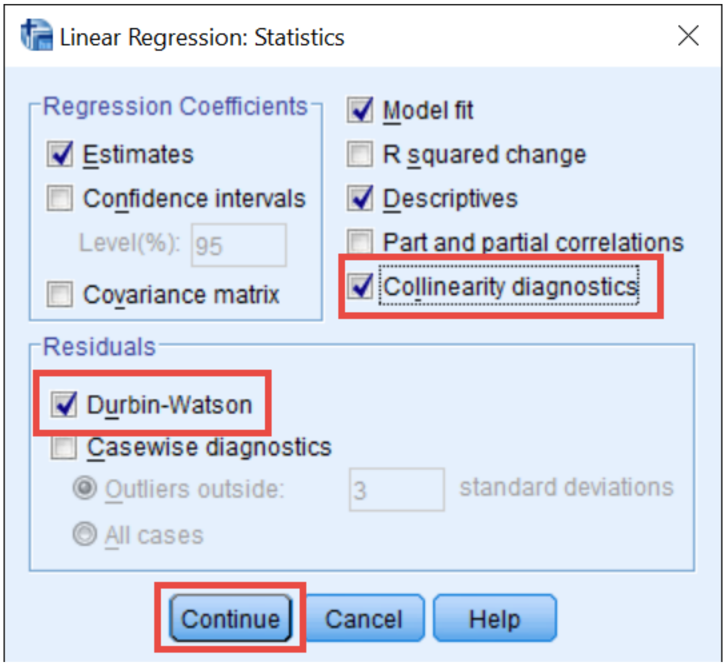
WLS Weight:

This opens the main Regression dialog box. As we haven't shut SPSS down since running our multiple regression (in the previous tutorial), SPSS remembers the options we chose for running our analysis. Exam Score is still selected as our DV, and Revision Intensity and Subject Enjoyment are entered as the predictors (or IVs).

To test the next couple of assumptions, **CLICK** on the **Statistics** option now.

Assumption #2: There is no multicollinearity in your data. This is essentially the assumption that your predictors are not too highly correlated with one another. To test this assumption, **SELECT Collinearity diagnostics.**

Assumption #3: The values of the residuals are independent. This is basically the same as saying that we need our observations (or individual data points) to be independent from one another (or uncorrelated). We can test this assumption using the **Durbin-Watson** statistic, so **SELECT** this option.



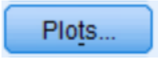
Regression Coefficients
☑ Estimates
□ Confidence intervals
Level(%): 95
□ Covariance matrix

☑ Model fit
□ R squared change
☑ Descriptives
□ Part and partial correlations
☑ Collinearity diagnostics

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

CLICK Continue to continue.

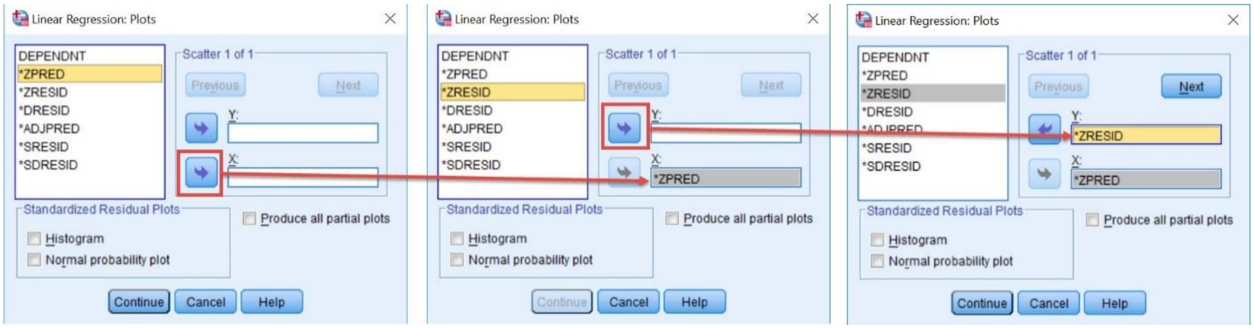
To test the next assumption, **CLICK** on the **Plots** option in the main Regression Dialog box.



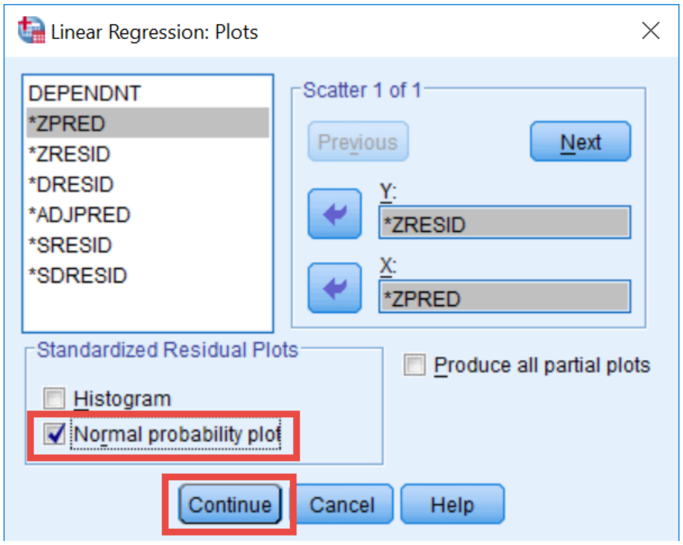
Assumption #4: The variance of the residuals is constant.

This is called *homoscedasticity*, and is the assumption that the variation in the residuals (or amount of error in the model) is similar at each point across the model. In other words, the spread of the residuals should be fairly constant at each point of the predictor variables (or across the linear model). We can get an idea of this by looking at our original scatterplot… but to properly test this, we need to ask SPSS to produce a special scatterplot for us that includes the whole model (and not just the individual predictors).

To test the 4th assumption, we need to plot the standardised values our model would predict, against the standardised residuals obtained.



To do this, first **CLICK** on the **ZPRED** variable and **MOVE** it across to the **X-axis.** Next, **SELECT** the **ZRESID** variable and **MOVE** it across to the **Y-axis.**



Y: *ZRESID

X: *ZPRED

□ Histogram

☑ Normal probability plot

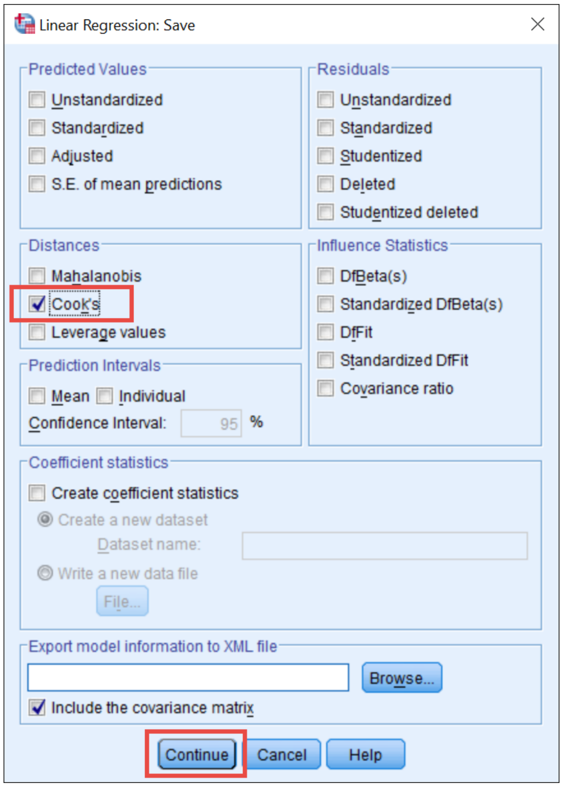
Assumption #5: The values of the residuals are normally distributed. This assumption can be tested by looking at the distribution of residuals. We can do this by **CHECKING** the **Normal probability plot** option.

Next, **SELECT Continue**

This brings us back to the main Regression dialog box.

Assumption #6: There are no influential cases biasing your model.

Significant outliers and influential data points can place undue influence on your model, making it less representative of your data as a whole. To identify any particularly influential data points, first **<u>CLICK</u>** the **SAVE** option in the main Regression dialog box.


Linear Regression: Save

Predicted Values
- ☐ Unstandardized
- ☐ Standardized
- ☐ Adjusted
- ☐ S.E. of mean predictions

Residuals
- ☐ Unstandardized
- ☐ Standardized
- ☐ Studentized
- ☐ Deleted
- ☐ Studentized deleted

Distances
- ☐ Mahalanobis
- ☑ Cook's
- ☐ Leverage values

Influence Statistics
- ☐ DfBeta(s)
- ☐ Standardized DfBeta(s)
- ☐ DfFit
- ☐ Standardized DfFit
- ☐ Covariance ratio

Prediction Intervals
- ☐ Mean ☐ Individual
- Confidence Interval: 95 %

Coefficient statistics
- ☐ Create coefficient statistics
- ● Create a new dataset
- Dataset name:
- ○ Write a new data file

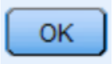
Export model information to XML file
- ☑ Include the covariance matrix

You can test for influential cases using **Cook's Distance**.

<u>SELECT</u> the **Cook's** option now to do this.

Then **<u>CLICK</u>** on **Continue**

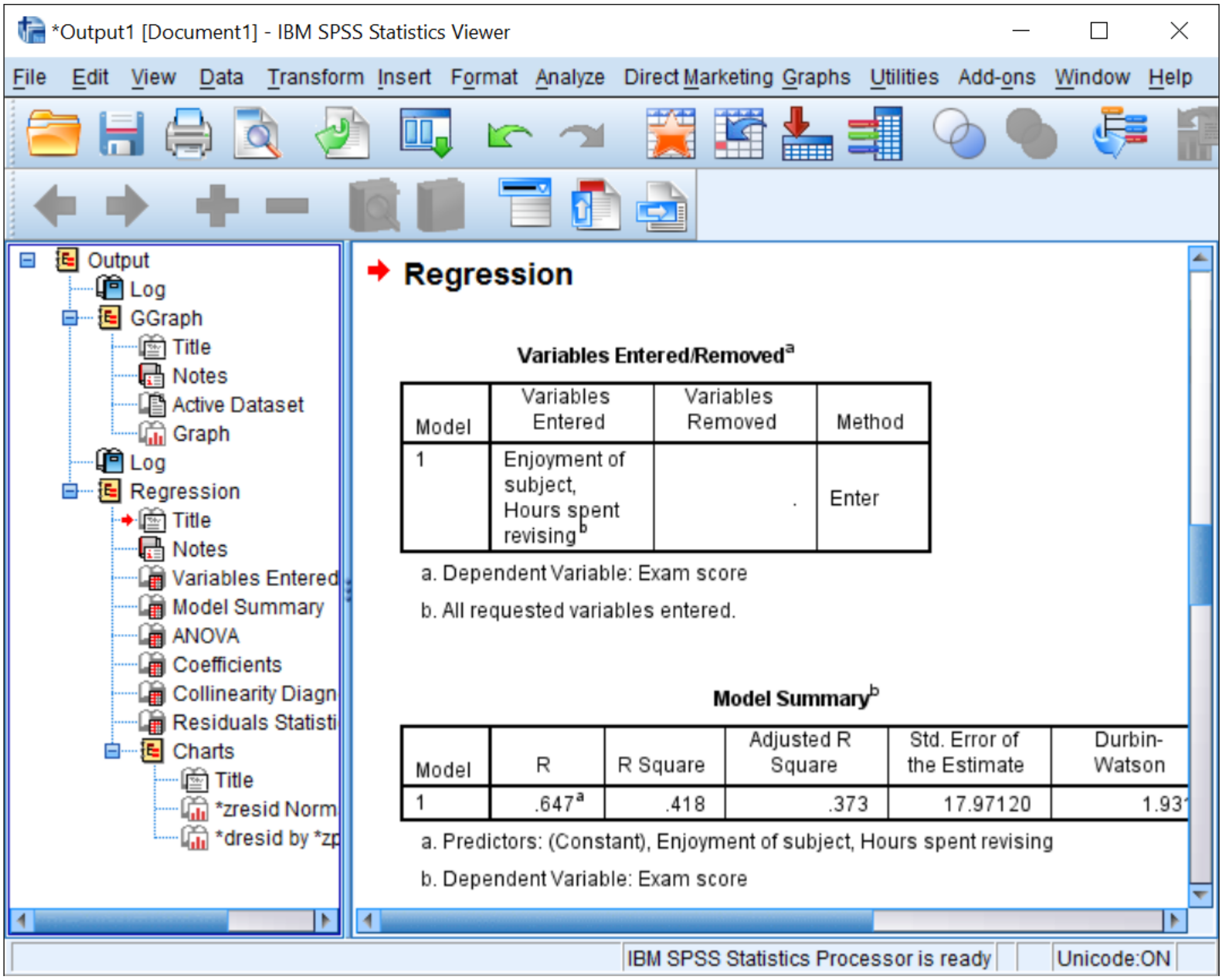
And finally **<u>CLICK</u>** on **OK** in the main Regression dialog box to run the analysis.



SPSS now produces both the results of the multiple regression, and the output for assumption testing. This tutorial will only go through the output that can help us assess whether or not the assumptions have been met.

To interpret the multiple regression, visit the previous tutorial.

The output appears in the SPSS Output window, below the scatterplot used to test Assumption #1. This tutorial will now take you through the SPSS output that tests the last 5 assumptions.



Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Enjoyment of subject, Hours spent revising[b]	.	Enter

a. Dependent Variable: Exam score

b. All requested variables entered.

Model Summary[b]

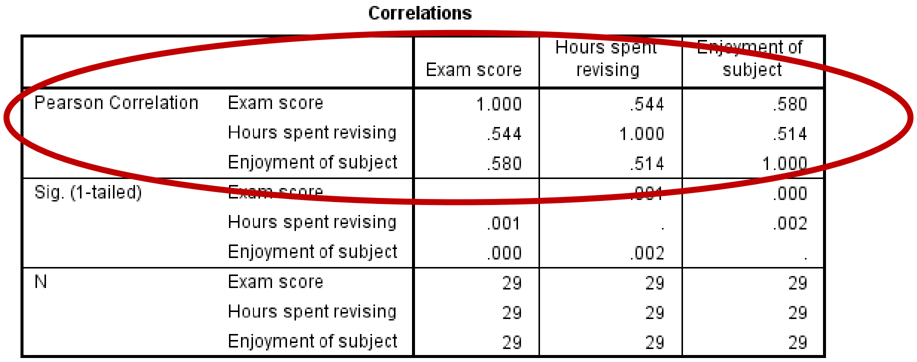
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.647[a]	.418	.373	17.97120	1.93

a. Predictors: (Constant), Enjoyment of subject, Hours spent revising

b. Dependent Variable: Exam score

Assumption #2: There is no multicollinearity in your data.

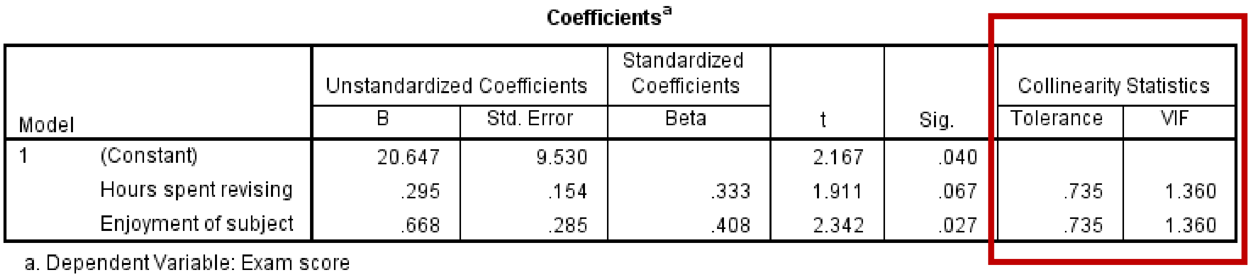
The first assumption we can test is that the predictors (or IVs) are not too highly correlated. We can do this in two ways. First, we need to look at the Correlations table. Correlations of more than 0.8 may be problematic. If this happens, consider removing one of your IVs. This is not an issue in this example, as the highest correlation is r=.58.



Correlations

		Exam score	Hours spent revising	Enjoyment of subject
Pearson Correlation	Exam score	1.000	.544	.580
	Hours spent revising	.544	1.000	.514
	Enjoyment of subject	.580	.514	1.000
Sig. (1-tailed)	Exam score		.001	.000
	Hours spent revising	.001	.	.002
	Enjoyment of subject	.000	.002	.
N	Exam score	29	29	29
	Hours spent revising	29	29	29
	Enjoyment of subject	29	29	29

We can also test this assumption by looking at the **Coefficients** table (note – you will have to scroll down the Output to find this table). This allows us to more formally check that our predictors (or IVs) are not too highly correlated. We can use VIF and Tolerance statistics to assess this assumption.



Coefficients[a]

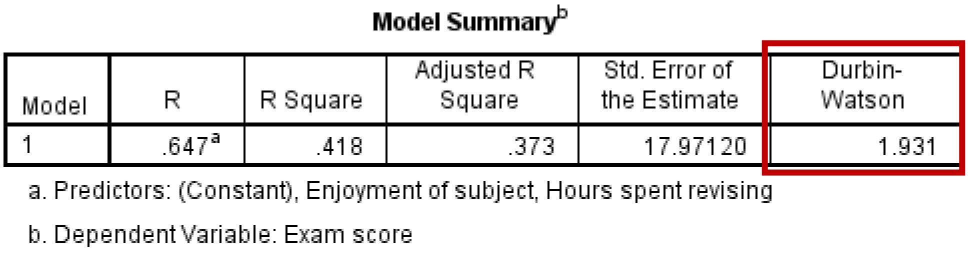
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	20.647	9.530		2.167	.040		
	Hours spent revising	.295	.154	.333	1.911	.067	.735	1.360
	Enjoyment of subject	.668	.285	.408	2.342	.027	.735	1.360

a. Dependent Variable: Exam score

For the assumption to be met we want VIF scores to be well below 10, and tolerance scores to be above 0.2; which is the case in this example.

Assumption #3: The values of the residuals are independent.

To check the next assumption we need to look at is the **Model Summary** box. Here, we can use the Durbin-Watson statistic to test the assumption that our residuals are independent (or uncorrelated). This statistic can vary from 0 to 4. For assumption #3 to be met, we want this value to be close to 2. Values below 1 and above 3 are cause for concern and may render your analysis invalid.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.647[a]	.418	.373	17.97120	1.931

a. Predictors: (Constant), Enjoyment of subject, Hours spent revising

b. Dependent Variable: Exam score

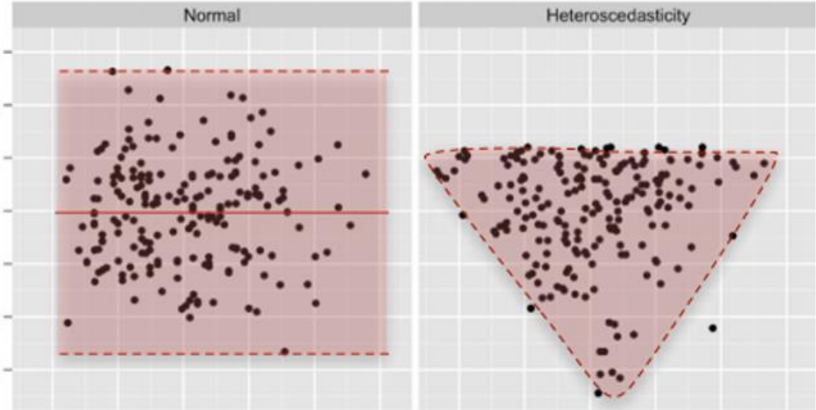
In this case, the value is 1.931, so we can say this assumption has been met.

Assumption #4: The variance of the residuals is constant.

To test the fourth assumption, you need to look at the final graph of the output. This tests the assumption of *homoscedasticity*, which is the assumption that the variation in the residuals (or amount of error in the model) is similar at each point of the model.

This graph plots the standardised values our model would predict, against the standardised residuals obtained. As the predicted values increase (along the X-axis), the variation in the residuals should be roughly similar. If everything is ok, this should look like a random array of dots. If the graph looks like a funnel shape, then it is likely that this assumption has been violated.

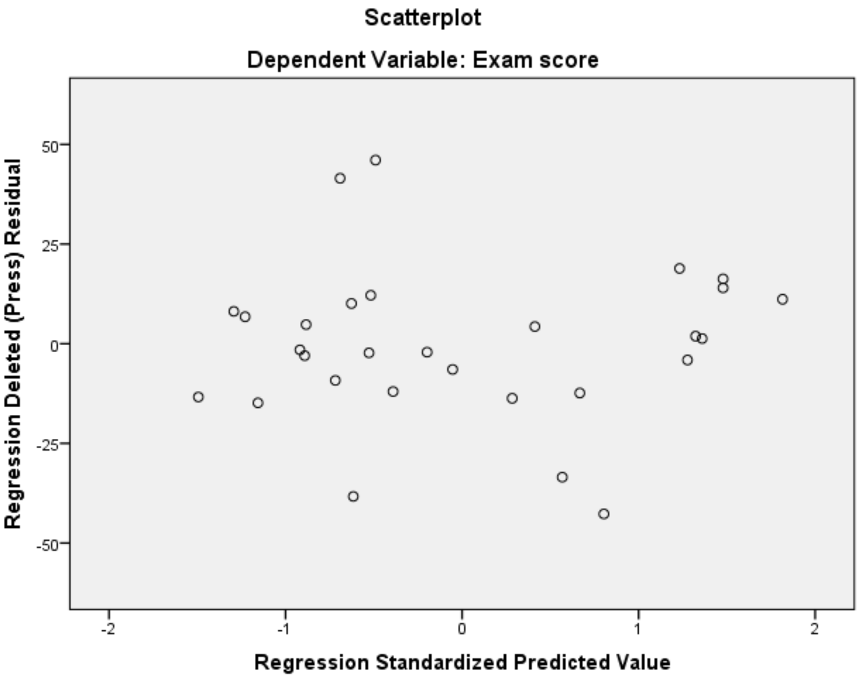
Assumption Met



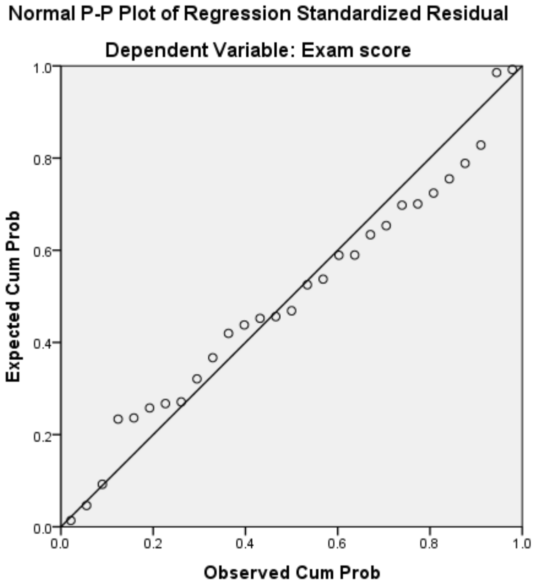
Assumption Not Met

Let's look at our scatterplot:

As we only have a small number of data points in this example, the graph can be difficult to read - but as it generally appears more random than funnelled, this assumption is probably ok. Again, any violations need to be reported in the write up.



Assumption #5: The values of the residuals are normally distributed.



This assumption can be tested by looking at the **P-P plot** for the model. The closer the dots lie to the diagonal line, the closer to normal the residuals are distributed.

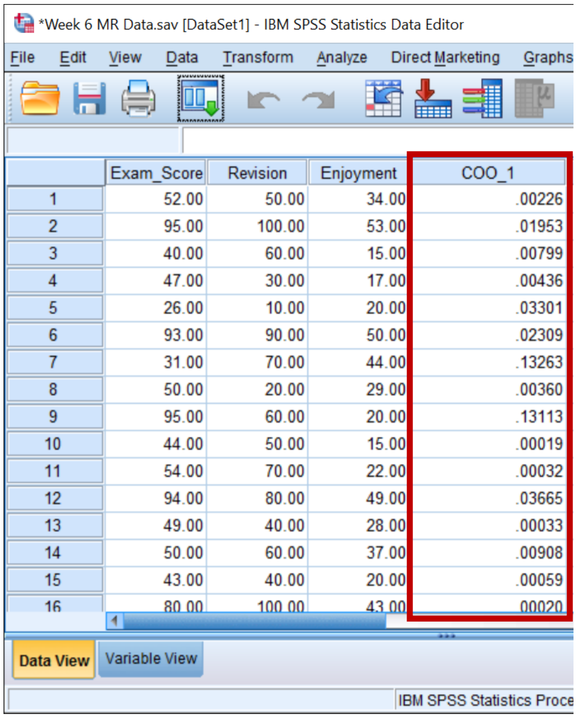
In this case, our data points hardly touch the line at all, indicating that assumption #5 may be violated. This will need to be flagged when writing up the results of the analysis, to let the reader know that they should be interpreted with caution.

Assumption #6: There are no influential cases biasing your model.

Our final assumption can be tested by going back to our Data File and looking at the Cook's Distance values we told SPSS to save for us. You can see SPSS has created a new column in our data file.

This contains the Cook's Distance statistic for each participant. Any values over 1 are likely to be significant outliers, which may place undue influence on the model, and should therefore be removed and your analysis rerun.

In this case, no such instances have occurred.


*Week 6 MR Data.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Data View
Variable View
IBM SPSS Statistics Proce

	Exam_Score	Revision	Enjoyment	COO_1
1	52.00	50.00	34.00	.00226
2	95.00	100.00	53.00	.01953
3	40.00	60.00	15.00	.00799
4	47.00	30.00	17.00	.00436
5	26.00	10.00	20.00	.03301
6	93.00	90.00	50.00	.02309
7	31.00	70.00	44.00	.13263
8	50.00	20.00	29.00	.00360
9	95.00	60.00	20.00	.13113
10	44.00	50.00	15.00	.00019
11	54.00	70.00	22.00	.00032
12	94.00	80.00	49.00	.03665
13	49.00	40.00	28.00	.00033
14	50.00	60.00	37.00	.00908
15	43.00	40.00	20.00	.00059
16	80.00	100.00	43.00	.00020

How to Write it Up

It is important that you flag any violations of your assumptions when writing up the results of your multiple regression analysis. In this case:

- **Assumption #1: The relationship between the IVs and the DV is linear**. ✓
  Scatterplots show that this assumption had been met (although you would need to formally test each IV yourself).

- **Assumption #2: There is no multicollinearity in your data.** ✓
  Analysis of collinearity statistics show this assumption has been met, as VIF scores were well below 10, and tolerance scores above 0.2 (statistics = 1.36 and .74 respectively).

- **Assumption #3: The values of the residuals are independent.** ✓
  The Durbin-Watson statistic showed that this assumption had been met, as the obtained value was close to 2 (Durbin-Watson = 1.93).

- **Assumption #4: The variance of the residuals is constant.** ✓
  Our plot of standardised residuals vs standardised predicted values showed no obvious signs of funnelling, suggesting the assumption of homoscedasticity has been met.

- **Assumption #5: The values of the residuals are normally distributed.** ✗
  The P-P plot 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 have a significant impact on your findings, the results are probably still valid.

- **Assumption #6: There are no influential cases biasing your model**. ✓
  Cook's Distance values were all under 1, suggesting individual cases were not unduly influencing the model.

This brings us to the end of the tutorial.

You have now seen how to test the assumptions of multiple regression using SPSS. Why not download the data file used in this example, and try to produce the output yourself. Remember, it is important to report any violations of these assumptions when writing up your results, so readers know that they should be interpreted with caution.