Research Methods

Linear Regression

As a researcher you are interested in working out how well mock exam marks can predict final exam marks. You decide to collect data from 15 participants (as shown below).

Because you are aiming to **predict** scores, we are looking at the analysis of Linear Regression (rather than correlation).

Step 1)

Create your two variables in the 'variable view' and input the data shown below in the 'data view'.

	MockExam	FinalExam
1	52.00	53.00
2	67.00	66.00
3	78.00	75.00
4	90.00	88.00
5	88.00	89.00
6	76.00	79.00
7	56.00	56.00
8	74.00	80.00
9	50.00	55.00
10	92.00	87.00
11	83.00	80.00
12	88.00	85.00
13	76.00	76.00
14	89.00	91.00
15	99.00	96.00
16		

Step 2)

Create a scatterplot for this data to visually determine the direction of the prediction.

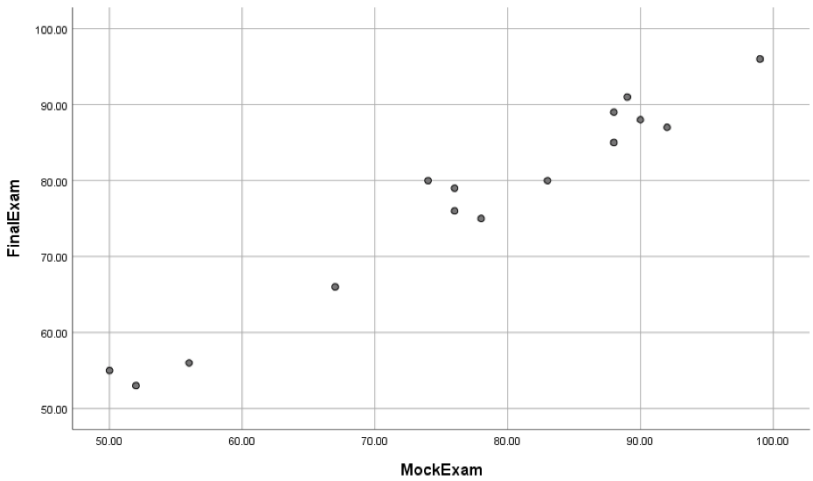
On the top menu bar, select 'Graphs', hover over 'Legacy Dialogs', and then scroll down to select 'Scatter/Dot'. Then choose 'Simple Scatter' and click Define.

Transfer your **DV** (final exam score) to the Y axis box, and the IV (mock exam score) to the X axis box.

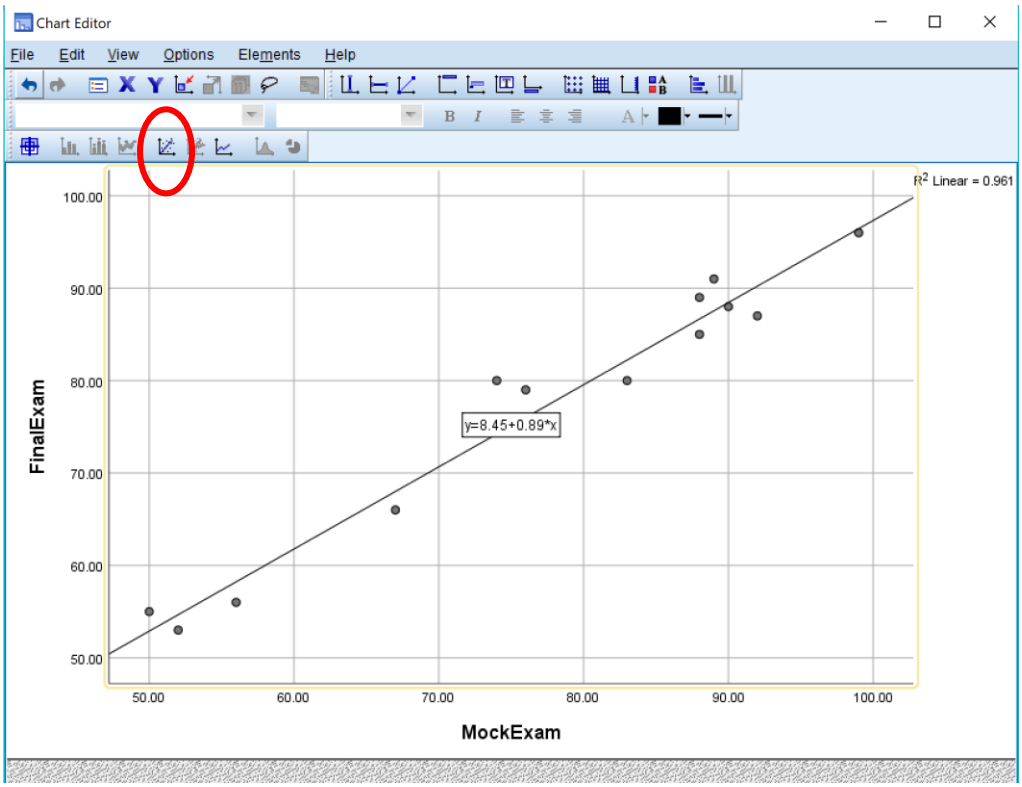
Note: You can click on 'Titles' if you wish to provide fuller descriptions of what is being displayed in the scatterplot.

Click OK to produce the scatterplot.

Now you will need to insert a 'line of best fit'. To do this, double click on the chart area to bring up the 'chart editor' window. Click on 'Add Fit Line at Total' (shown below). You can use this line to determine whether there looks to be a relationship between the two variables, and in what direction the relationship looks to be going in.



BUT! We do not yet know if this prediction is significant, so we need to conduct the statistical test to allow us to make this judgement.



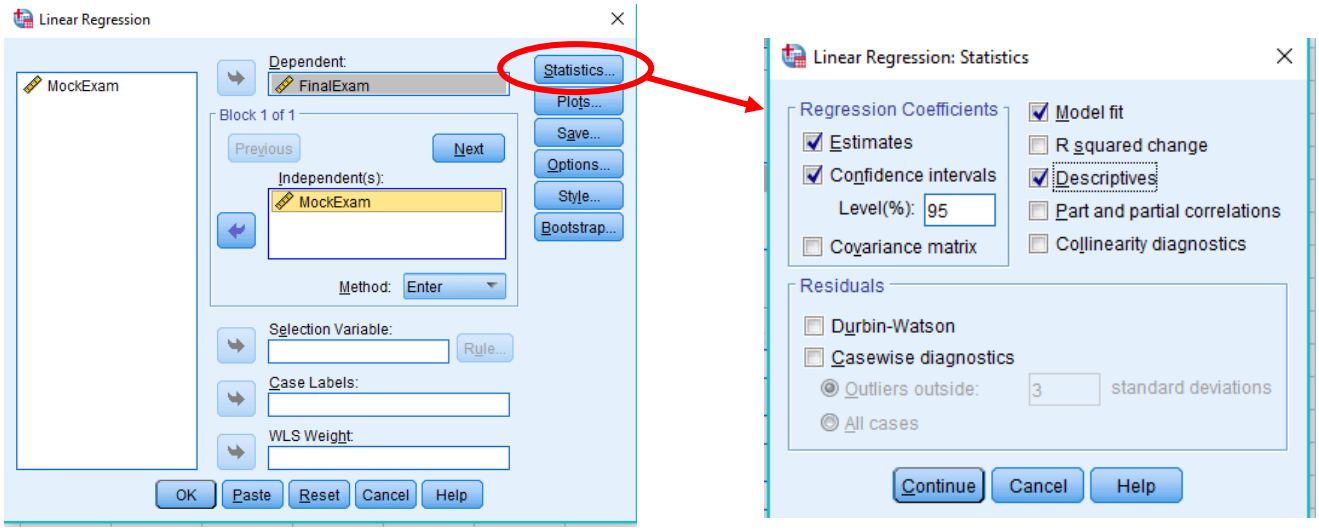
Step 3)

Conduct the Linear Regression analysis.

NB: there are two ways to do this – both procedures will produce the same outcome but, for this example, we will be running through the procedure that is more closely linked to the multiple regression analysis that you will be practicing in future workshops.

Analyze → Regression → Linear.

The linear regression window will appear and you should move your variables accordingly to the boxes shown below. Your DV in this case is Final Exam marks, and the IV is Mock Exam marks. Select 'Statistics' and select the options as shown below. Click 'Continue'.

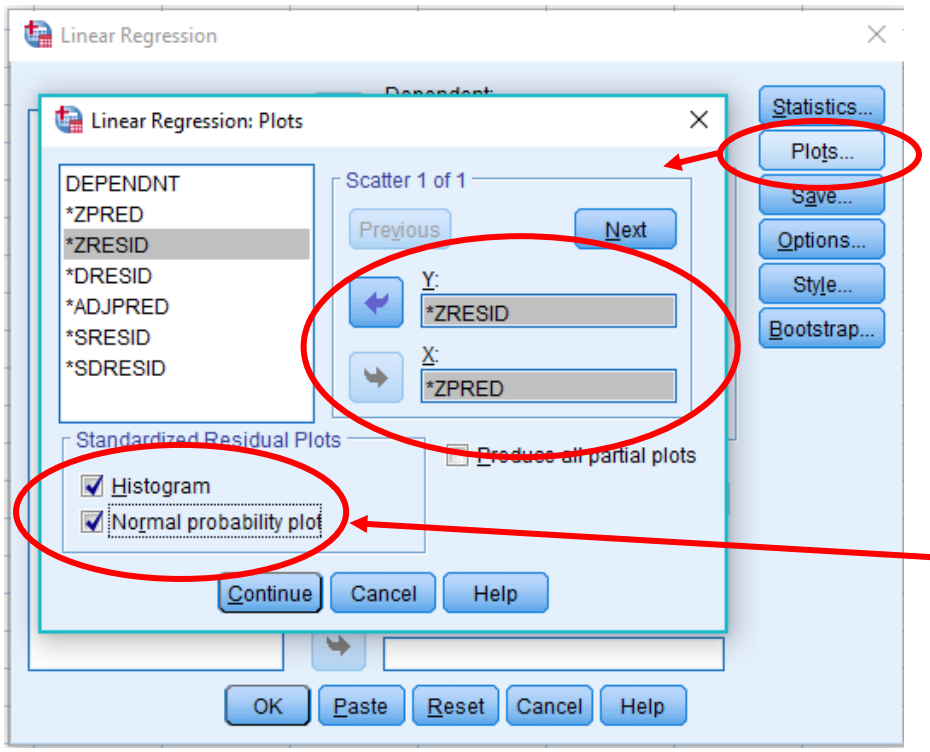


Step 4)

Next, click on Plots. Remember that there are assumptions regarding your data that need to be considered, and the plots option will help you to ensure these assumptions are met.

Move '*ZRESID' in to the Y box and move '*ZPRED' in to the X box.

Hopefully, the chart plotted from this will not show any pattern – so you can be sure that the amount of errors of similar for both variables. Click 'Continue' and then 'OK' to run the linear regression analysis.



You can also select these to obtain a histogram that will hopefully show a normal distribution for your DV. You should also obtain a normal P-P plot of regression standardised residuals, where hopefully you will see that all the data points are reasonably close to the line. If this is the case, then you can assume normality.

SPSS Output

Descriptive Statistics: shows the average scores for each of the variables included in the analysis.

Correlations: this table will provide you with the r-value (correlation) between the predictor and criterion variables, and the associated levels of significance.

Model Summary:

You will see here that the value under 'R' is the same as the correlation value in the correlations table.

The value you are most interested here is the 'R Square': this value indicates how much of the total variation in the DV can be explained by the IV. We need to transfer the value in this box to a percentage (%).

ANOVA:

The information contained within this table allows us to determine how well the regression equation fits the data (i.e., predicts the DV). The information from this table gets reported as a typical ANOVA.

Coefficients:

This table provides us with the necessary information to allow us to predict DV scores, as well as to determine whether our IV significantly contributes to the model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	8.452	3.888		2.174	.049	.053	16.852
	MockExam	.889	.049	.980	17.961	.000	.782	.996

a. Dependent Variable: FinalExam

This is *a* (the intercept)

This is *b* (the slope)

This has converted *b* into a standardised regression coefficient, which is measured in units of standard deviation

The smaller the range between these two numbers, the more confident we can be in our findings. These values tell us the range in which the value of the population slope can be found

Writing up the results of the linear regression.

> "A simple linear regression was carried out to determine whether finals marks on an exam could be predicted by marks scored on a mock exam. A significant regression equation was found (F (1, 13) = 322.58, $p < .001$). The adjusted R^2 was 0.96, suggesting that 96% of the variation in the final exam marks can be explained by the mock exam marks. For every one mark increase in the mock exam scores, final exam marks increased by .98. The range between confidence limits were narrow, highlighting that it is 95% likely that the population slope is between 0.78 and 1.00.

Predicting new cases.

You can use the information provided in the regression output to predict new cases in the regression equation.

Y = final exam mark
X = mock exam mark
a = 8.45
b = 0.89

$$\mathbf{Y = bX + a}$$

Try using the equation above and input a pretend mock exam mark in X to calculate the equation and see what result you get. This would be the individuals predicted final exam mark.