

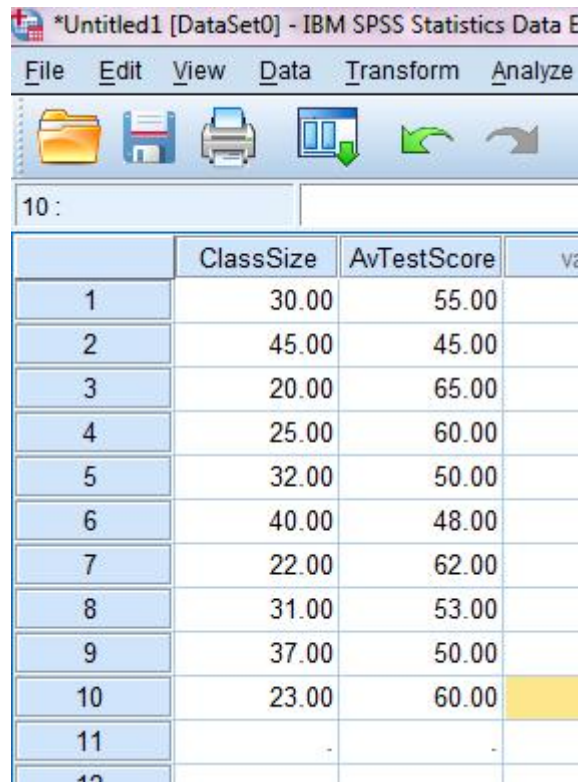
Research Design and Analysis

Correlation and Linear Regression

In this workshop you will practice conducting Correlation and Linear Regression analyses using SPSS. Please use the information provided in this worksheet to complete the examples during the workshop.

Correlation

1. An educational psychologist is interested in exploring whether class size relates to average test scores from 10 school children. The data from this study are shown below.



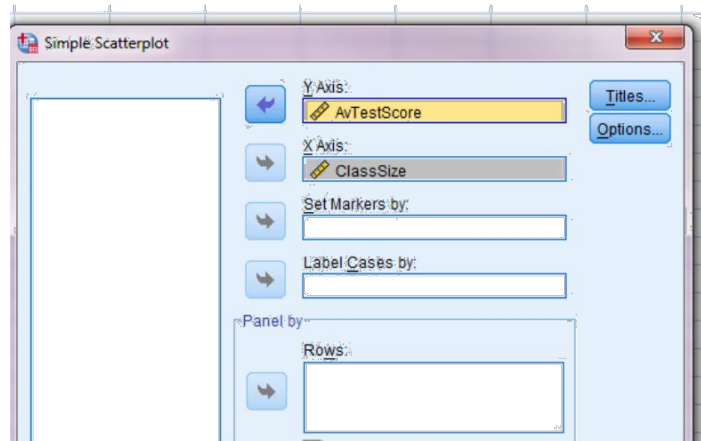
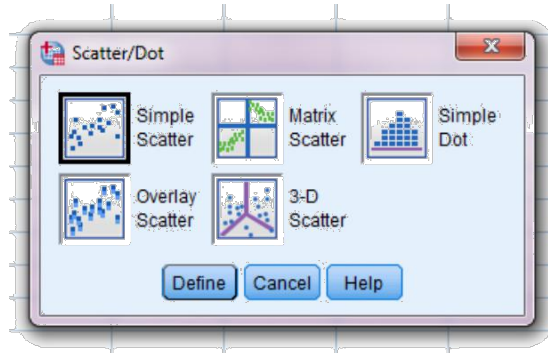
	ClassSize	AvTestScore	vs
1	30.00	55.00	
2	45.00	45.00	
3	20.00	65.00	
4	25.00	60.00	
5	32.00	50.00	
6	40.00	48.00	
7	22.00	62.00	
8	31.00	53.00	
9	37.00	50.00	
10	23.00	60.00	
11			
12			

Name your variables as appropriate in the 'variable view', and input the data shown above using 'data view'.

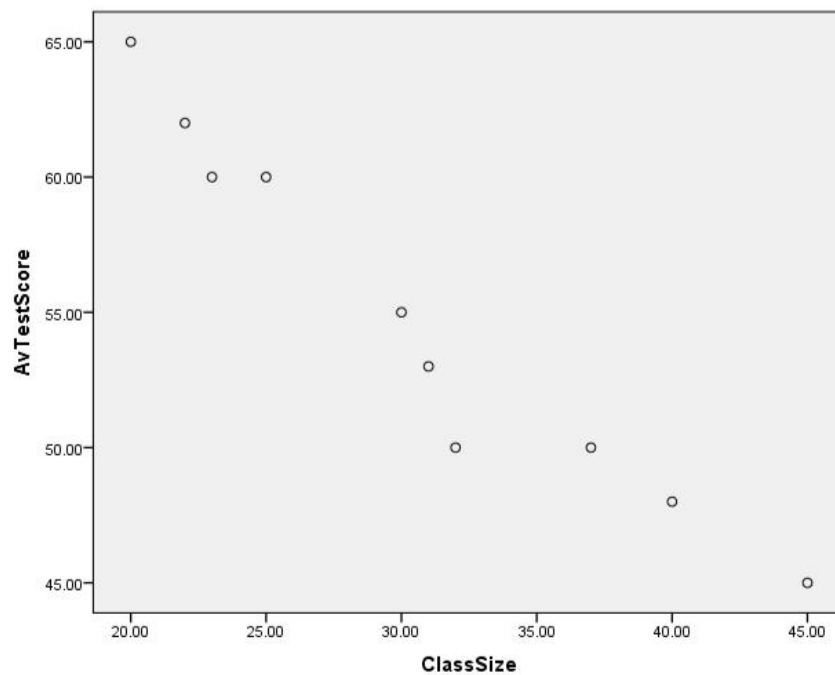
The first thing to do with a correlation is to create a scatterplot diagram to visually see the direction of the correlation. On the top menu bar, select 'Graphs', hover over 'Legacy Dialogs' and then scroll down to select 'Scatter/Dot'. Then choose 'Simple Scatter' and click Define.

Transfer your **DV** (test score) to the 'Y Axis' box, and the IV (class size) 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.

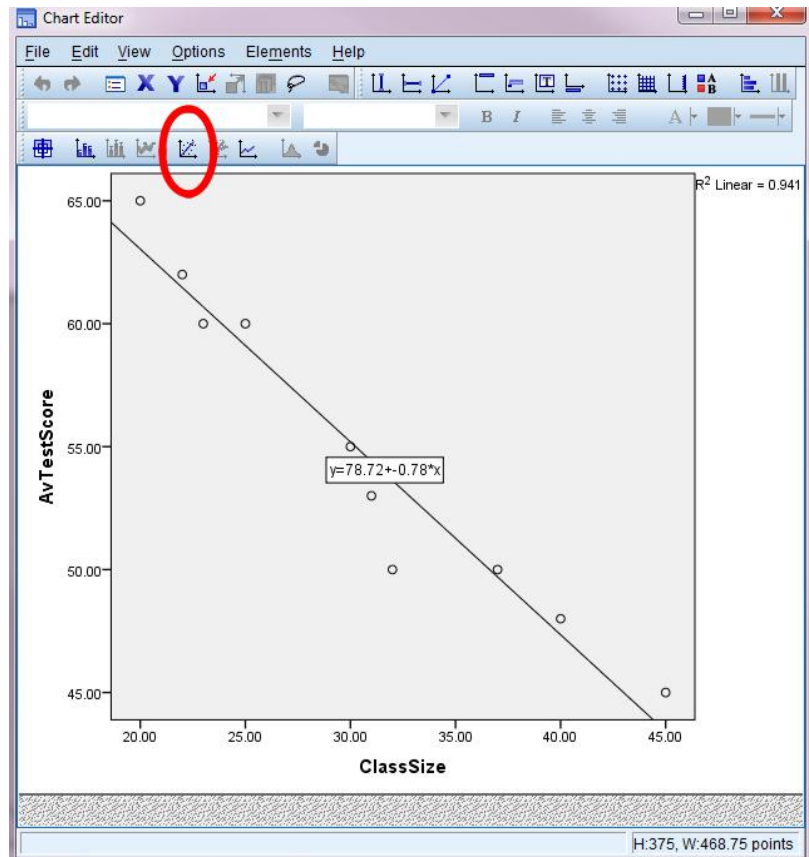


Click 'OK' to produce the scatterplot of your data.



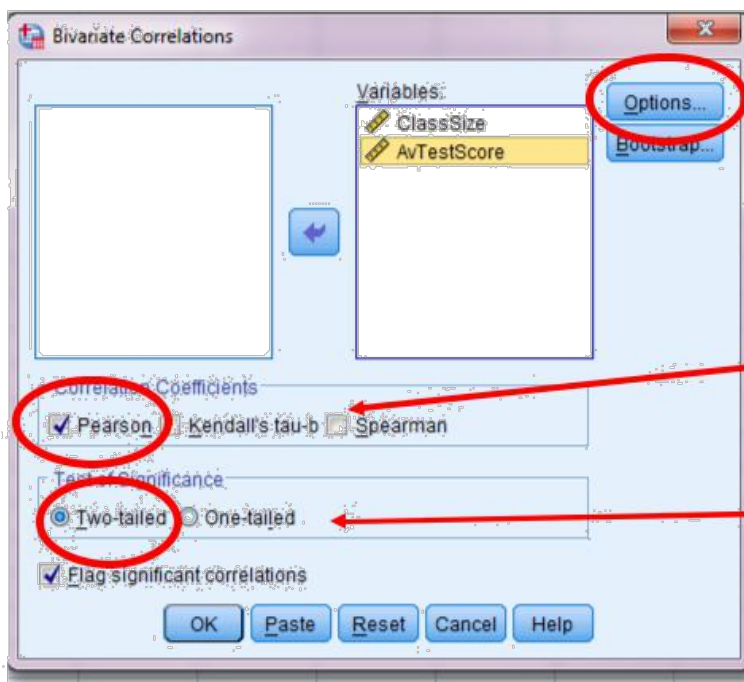
Now you will need to insert a 'line of best fit'. To do this, double click on the chart area. A 'chart editor' window will appear. Click on 'Add Fit Line at Total' (shown below). You can use this line to determine whether there is a correlation between our two variables, and what direction this relationship goes (positive or negative).

BUT! We do not yet know if the correlation is significant; so we need to conduct the statistical test to make this judgement.



To run the correlation, select 'Analyze' on the top menu bar, then scroll down to 'Correlate' and click on 'Bivariate'.

In the next window, move both of your variables in to the 'Variables' box. Click OK to run the analysis when you are sure everything has been entered correctly.



Click on 'Options' to include 'Means' and 'Standard Deviations' in your output

Choose which statistic you wish to perform – Pearson's, in this case

You can also specify whether the test is one-tailed or two-tailed

The output for this analysis will provide you with two boxes.

Descriptive Statistics will show you the mean and SD scores for each of your variables.

The **Correlations** table will provide you with the information you require to determine whether the relationship between the variables is significant or not.

		ClassSize	AvTestScore
ClassSize	Pearson Correlation	1	-.970**
	Sig. (2-tailed)		.000
	N	10	10
AvTestScore	Pearson Correlation	-.970**	1
	Sig. (2-tailed)	.000	
	N	10	10

** . Correlation is significant at the 0.01 level (2-tailed).

The correlation value comes from here

Note: The minus sign tells you that there is a **negative** correlation – NOT that there is NO correlation

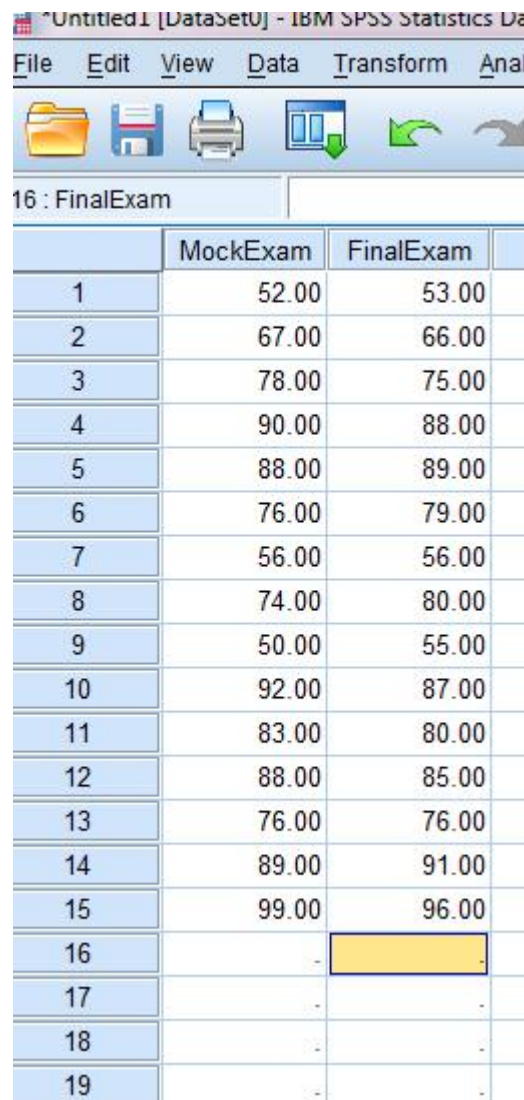
The significance value comes from here

The Pearson Correlation value (r) indicates the strength of the correlation. The Sig. value will tell you if the two variables are significantly related. Write the results of this correlation below:

Linear Regression

2. 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).

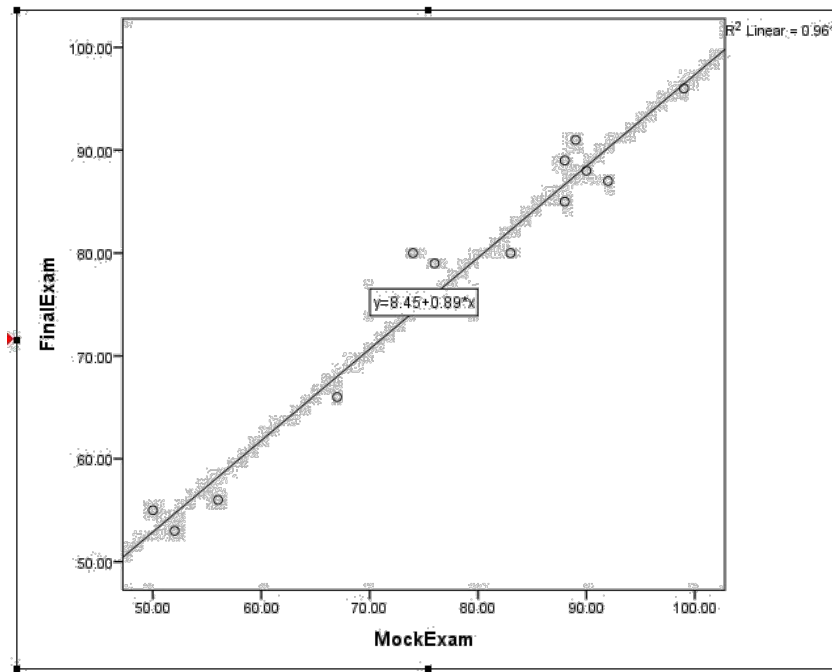
Create your two variables in the 'variable view' and input the data below in the 'data view'.



The screenshot shows the IBM SPSS Statistics Data Editor window. The title bar reads "Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor". The menu bar includes File, Edit, View, Data, Transform, and Analyze. The toolbar contains icons for opening a file, saving, printing, and other standard functions. The variable view at the top shows two variables: "MockExam" and "FinalExam", both of type "Scale". The data view below shows 15 rows of data. The first column is labeled "16 : FinalExam" in the header, but the data is organized by row number. The columns are "MockExam" and "FinalExam". The data for rows 1 through 15 is as follows:

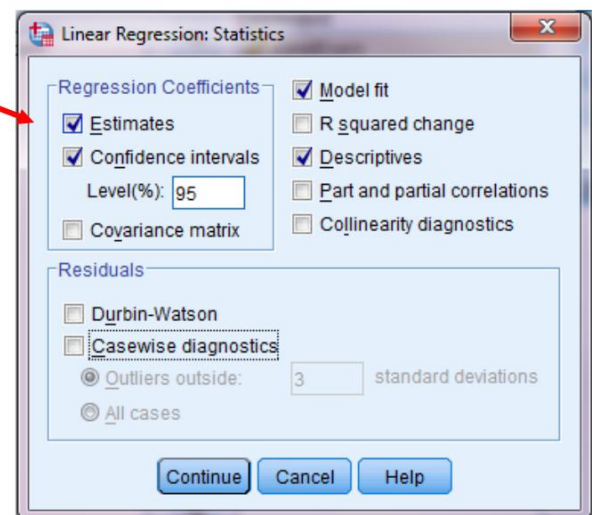
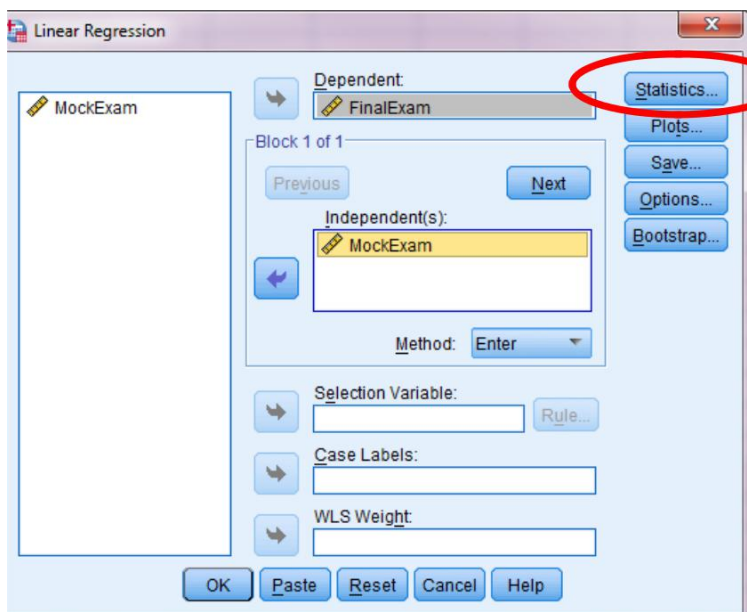
	MockExam	FinalExam
1	52.00	53.00
2	67.00	66.00
3	78.00	75.00
4	90.00	88.00
5	88.00	89.00
6	76.00	79.00
7	56.00	56.00
8	74.00	80.00
9	50.00	55.00
10	92.00	87.00
11	83.00	80.00
12	88.00	85.00
13	76.00	76.00
14	89.00	91.00
15	99.00	96.00
16	-	-
17	-	-
18	-	-
19	-	-

Create a scatterplot for this data using the method outlined in the previous section. You will end up with something like this:

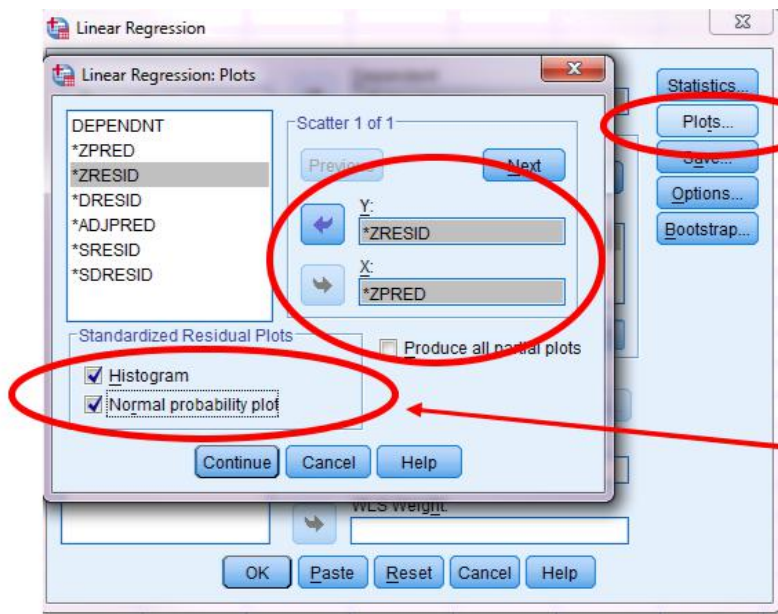


Now, we want to conduct the Linear Regression. There are two ways to do this – both procedures will produce the same outcome. For this workshop, we will be running through one of these (the one that will be more useful for you to understand in preparation for your multiple regression workshops in the upcoming weeks).

Go to 'Analyze' on the top menu, hover over 'Regression' and select 'Linear'. The Linear Regression window will appear and you will need to 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 tickboxes as shown below. Click 'Continue'.



Next, clicks 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 residual, where hopefully you will see all the data points reasonably close to the line. If this is the case, then you can assume normality.

The output provided will look very different from the correlation output above.

Descriptive Statistics

	Mean	Std. Deviation	N
FinalExam	77.0667	13.71895	15
MockExam	77.2000	15.13369	15

Correlations

		FinalExam	MockExam
Pearson Correlation	FinalExam	1.000	.980
	MockExam	.980	1.000
Sig. (1-tailed)	FinalExam	.	.000
	MockExam	.000	.
N	FinalExam	15	15
	MockExam	15	15

Correlation between the predictor and criterion variables and the associated levels of significance

IMPORTANT!!

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.980 ^a	.961	.958	2.80212

a. Predictors: (Constant), MockExam

b. Dependent Variable: FinalExam

The correlation
(same value as
box above)

Takes account of No. of Ps &
variables in the model. Give a
more realistic estimate of
amount of variance explained.
Need to transfer to %

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2532.859	1	2532.859	322.581	.000 ^b
	Residual	102.074	13	7.852		
	Total	2634.933	14			

a. Dependent Variable: FinalExam

b. Predictors: (Constant), MockExam

Associated degrees of
freedom values

Shows whether the
regression line is significantly
different from 0 (note that
this relates more to multiple
regression – we only have
one variable in this model so
we could just use the
correlation significance.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	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
in to 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

$$Y = bX + a$$

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.