

Q-Step Training Workshop

Data Analysis (intermediate level)

(practical instructions.. see accompanying slides)

Mark Brown

Step 1: OPEN THE DATASET

1. **First Open SPSS** (go to Start... type SPSS into search box and select 'IBM SPSS Statistics 22' to open the software)
2. **Then** open the dataset ('World Bank 2017') .. do this from menu in SPSS i.e. select File.. Open..etc)

For these practical exercises we'll focus on the **U5 Mortality rate (deaths per 1000)** as our chosen measure of interest (the dependent variable) but you can repeat the exercises later with your own choice

Step 2: ANALYSING THE DATA

There are 4 sections to this analysis:

1. **Always start by looking at the Univariates**
2. **Exploring relationships graphically with a Scatterplot**
3. **Measuring Correlation**
4. **Simple Regression**

1. ALWAYS START BY LOOKING AT THE UNIVARIATES

Use appropriate techniques to show the distribution of the **U5 Mortality rate** (could use a graph, summary statistics or both)

Q> Can you use the outputs to summarise the global distribution of U5 Mortality rates?

2. EXPLORING RELATIONSHIPS GRAPHICALLY WITH A SCATTERPLOT

a) **Produce a scatterplot** to show the relationship between **U5Mortality rate** and **% of women in secondary education**

Scatterplots can be produced by selecting from the menu...

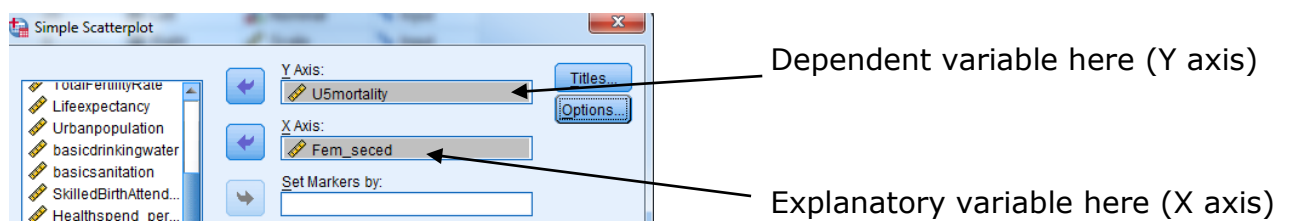
Graphs...

Legacy Dialogues...

Scatter/dot...

Simple scatter

Note: where you are able to identify which variable is dependent and which is explanatory you should make the **dependent variable the y axis variable**



Click ok to produce the scatterplot

Q> Does the scatterplot suggest a relationship? If so try and summarise it

b) **Run another scatterplot** – this time use an explanatory variable of your own choice (a measure you think will be related to U5mortality)

3. MEASURING CORRELATION

a) **Compute Pearson correlations**

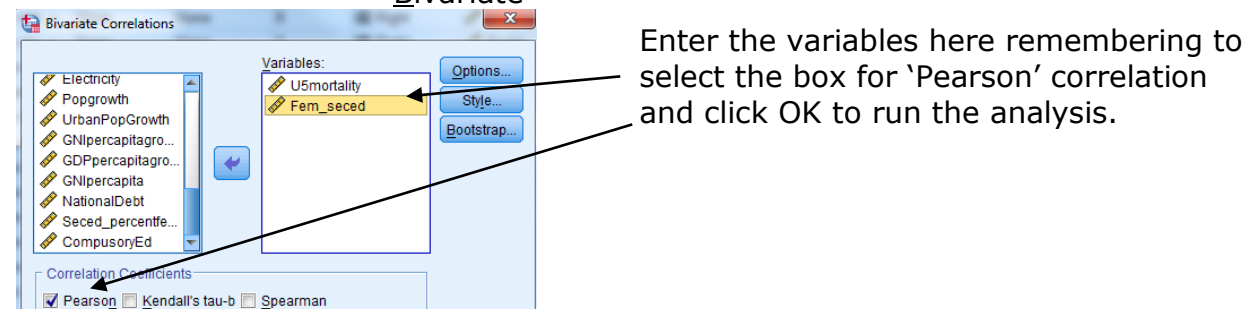
Compute a Pearson correlation for U5 mortality and %women in secondary

The correlation procedure can be selected from the menu...

Analyze...

Correlate...

Bivariate



This will produce a **'correlation matrix'**

Q> can you interpret the correlation value? Is it statistically significant? (tip: check the p values)

Now repeat the correlation this time between U5 mortality and the other explanatory variable you used in the scatter plots

Q> Which explanatory variable has the strongest correlation with U5Mortality?

Q> Why should you be cautious in assuming a causal link between the explanatory and dependent variable?

b) Compute Spearman Rank correlations

Repeat the steps for the Pearson correlation (but click on 'Spearman' rather than 'Pearson'). You'll probably get similar results but note the following..

Which to use, Pearson or Spearman?

For Pearson correlation

- b) Both variables must be interval level and ideally normally distributed
- c) It is only suitable if the relationship is linear (straight line). It will not be sensitive to relationships that are non-linear (so always check with a scatterplot first)

Spearman Rank Correlation may be used in place of Pearson Correlation in the following two situations.

- a) When one or more of the variables is ordinal (not relevant in our case) or deviates from a normal distribution
 - b) When the relationship is non-linear (may be the case – check your scatterplots)
- Sometimes a scatterplot between two interval variables shows a relationship that is curved (curvilinear) rather than strictly linear. In these cases Spearman's correlation is a more appropriate measure of correlation.

4. SIMPLE REGRESSION

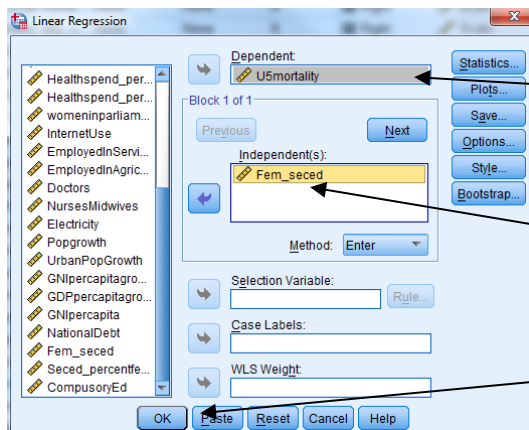
Simple regression fits a straight line to describe the relationship between a dependent variable and an independent (explanatory) variable. The line is described by a simple equation which you can then use to predict values of the dependent variable for given values of the independent variable.

Linear regression is only a valid technique where the relationship between the two variables is linear. So you always need to check the scatterplots and Pearson correlation results first.

a) Running a Simple Regression Model

To run the regression, from the menu select...

Analyze... Regression... Linear



Dependent variable goes here (U5mortality)

Independent (explanatory) variable goes here

Click ok to run the model.

Q> Use information in the results to fill in the gaps for the following regression equation.

Dependent variable = Constant (b_0) + (slope (b_1) x Explanatory variable)

U5Mortality rate = _____ + (_____ * % women with lower sec education)

Q> Can you explain the meaning of the 'constant' and the 'slope' in the equation?

Q> Can you use the equation to predict the U5 mortality rate in a country with the following % of women in secondary education:

- a) 20% of women with lower sec education
- b) 80% of women with lower sec education

Q>. What percentage of the variation in U5mortality rate is explained by the explanatory variable? (what measure do you use to answer this)

Q>. Suggest why it is rare to get very high R^2 values using simple regression in social science research

b) Finally ... if you have time try running one or two more regression models using different explanatory variables to predict U5 Mortality – or try a different dependent variable.

MB June 2019