

# SPSS Lab 10

## Linear Regression and Correlation

### *Demo and Lab Assignment 1*

#### 1. Obtaining a Scatterplot

Open file countries\_birthdeathrates.sav. It contains birthrates, death rates, life expectancy, and other similar rates per country. Let's analyze if there is an association between birthrate and life expectancy.

- a. Go to Graphs\Legacy Dialogs\Scatter\Dot.... Select Simple Scatter. Click Define. Y axis=female life expectancy, x axis= birthrate, Label cases by: country. Set markers by: develop. Click Ok.
- b. Double Click the chart. Select Options/Show data labels. Well, maybe we have too many labels. Therefore, let's hide them. How? You can find out!
- c. Go to Elements\Fit Line at Total and click selection box "linear" Fit method.
- d. What does the line tell you?
- e. What does the R squared value tell you? ((If you are not taking BIL 311, google for "r squared regression"))

#### 2. Calculating the Least Squares Line and the correlation coefficient

- a. Write the hypothesis for this study
- b. Go to Analyze/Regression/Linear ...
- c. As the dependant variable choose "female life expectancy", as the independent variable choose " birth rate"
- d. Click Statistics and select Estimates, Confidence Intervals, Model fit, and descriptive
- e. Click on Plot and select DEPENDANT as the x variable and \*SRESID (the studentized residuals) as the Y variable and check "Produce all partial plots"
- f. Click on "Save" and select Predicted Values "Unstandardized" and Residuals "Unstandardized" and "Studentized." This will create three new field in your data set (PRE\_1,RES\_1,SRE\_1). Close that dialog. Click OK.
- g. Now let's analyze the results. Go to Model Summary. Look at the R and R squared values. What do they mean? Is one of them the same as the one in the scatter plot?
- h. Look at the ANOVA (analysis of variance) table. Does this look like something we discussed in class? If not (if you are not taking BIL 311) google for "ANOVA for linear regression." What is the main value of interest in that table and what does it mean?
- i. The coefficient table contains the slope and the intercept of the linear regression line in column B.
  - i. Which one is which?
  - ii. The values that you obtained for the slope and intercept are based on one sample from the population. If you take a different sample from the same population, you will get different values for the slope and intercept. The distribution of all possible values of the slope and intercept are normal if the regression assumptions are met. The standard deviations of these distributions are called the standard error of the slope,  $se(\text{slope})$  and the standard error of the intercept,  $se(\text{intercept})$ . They are estimated from the data. You will see

them under column Std. Error. What do you think about those values? (Small, large?)

- iii. We have another significance test here, this time it is a t-test. The t statistics is the slope/ se(slope). What is the t statistics here? What is the outcome of the test?
- j. The scatterplot of the residuals shows you the variability between observed and expected values. It is used to analyze the assumptions of the regression analysis. If the residuals fall between 2.5 and -2.5 the equal variance assumption is valid.
- k. Now let's look at the studentized residuals. The residuals should follow a normal distribution. Therefore let's analyze it with a Q-Q plot. Go to Analyze\Descriptive Statistics \Q-Q Plot. Select the studentized residual as the variable and hit OK. Do you think the residuals are normally distributed? Why again is this important to know?

*Lab Assignment*

- 3. Do 1 and 2 for one other combination of birth rate, death rate, female and male life expectancy rates. Report your results.