Practical "Correlation and simple linear regression"
P5

1. Dataset *armstrength.sav* contains data on lifetime alcohol intake (measured as kg/kg body weight) and the strength of the deltoid muscle in non-dominant arm (measured in kilograms) for 30 alcoholic men.

   (a) Make a scatter plot of arm strength (*armstrength*) vs *alcohol*. Would a linear fit be appropriate?
   NOTE: To make a scatter plot go to *Graphs* $\rightarrow$ *Legacy Dialogs* $\rightarrow$ *Scatter* and there select *Simple scatter*.

   (b) Estimate the coefficients for a simple linear regression model with arm strength as the dependent variable and as the independent variable. Check the assumptions for this model. What is the equation of the regression line? Interpret the slope of this line.
   NOTE: To run regression and to calculate residuals and predicted values go to *Analyze* $\rightarrow$ *Regression* $\rightarrow$ *Linear* and select the corresponding dependent and independent variables. Click on *Save*, select *Unstandardized* residuals and *Unstandardized* predicted values. Notice that two extra columns have been added to the data. Make a histogram of the unstandardized residuals (RES_1). To make a P–P plot go to *Analyze* $\rightarrow$ *Descriptive Statistics* $\rightarrow$ *P–P Plots* and there select unstandardized residuals as a variable (no need to change other options). To plot the residuals against *alcohol* go to *Graphs* $\rightarrow$ *Legacy Dialogs* $\rightarrow$ *Scatter/Dot (Simple)* with the residuals as the y-axis variable and *alcohol* as the x-axis variable. In a similar way make a plot of the residuals versus the predicted arm strength values (PRE_1).

   (c) At the 5% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that *alcohol* is useful as a predictor of arm strength?

   (d) The output from the linear regression model includes a table with *R* and *R–square*. Compare it to the Pearson correlation coefficient for the same variables. Do we reach the same conclusion as with a regression model? Interpret *R–square*.
   NOTE: To calculate the correlation coefficient go to *Analyze* $\rightarrow$ *Correlate* $\rightarrow$ *Bivariate*.

2. Dataset *usmelanoma.sav* contains male mortality rates of malignant skin cancer (expressed as number of deaths per 10 million people) for 49 states in the United States along with the degrees north latitude of a centroid of each state.

   (a) Create a scatter plot of *mortality* versus *latitude* using *latitude* as the explanatory variable and calculate the Pearson's correlation coefficient between these two variables. Use the results to conclude about the relationship between *mortality* and *latitude*.

   (b) Fit a linear regression line to the data, determine the regression equation and interpret the value of the slope coefficient.
   NOTE: To see the least-squares line, double click on the scatter plot to open Chart Editor, select all data points (they will become highlighted), then go to *Elements* $\rightarrow$ *Fit line at total* (change nothing) $\rightarrow$ Close the Chart Editor.

   Assuming that the data meet the assumptions of simple linear regression, answer the following questions:

(c) If we consider two states of which one has a latitude 5 degrees south of the other, how much difference is expected in the melanoma mortality rates?

(d) Using the regression line, predict the melanoma mortality rate for a state at the latitude of 40 degrees north. Is this prediction reliable? Provide explanation.
NOTE: You can calculate the predicted values using the regression line and a calculator. Alternatively, you can enter the number 40 in the *Latitude* variable column of the data window after the last row and enter a '.' for the corresponding *Mortality* variable value. Then, go to *Analyze* $\rightarrow$ *Regression* $\rightarrow$ *Linear*. Click *Save* and select *Unstandardized* predicted values. Go back to the data and check the predicted mortality value in the last row.

3. In exercise 6 of Practical P4 we looked at the dataset *btgdiabet.sav*, containing data on urinary beta-thromboglobulin (beta-TG) excretion in 12 normal subjects and in 12 diabetic patients (Kirkwood 2003). We did a log transformation to be able to use an independent samples *t*-test.

(a) Take the log transformed data and obtain results for the test again (or look at what you obtained).

(b) Compare the population means for both groups using simple linear regression instead. To do that, create dummy variables from *group* and estimate a simple linear regression model with one of these dummies, e.g. the one representing diabetic subjects (i.e. considering normal subjects as a reference group) as an explanatory variable. What values are exactly the same in the Coefficients table and in table generated at point a)? What do you conclude based on those findings?
NOTE: To create a dummy variable for group of normal subjects go to *Transform*$\rightarrow$ *Recode into different variables*. Then, choose *group* to be the numeric variable, fill in the name of the output variable and click *Change*. After that, click *Old and New Values* button and there under *Old Value* fill in value 1 and under *New Value* fill in value 1. Click *Add*. Then, under *Old Value* click on *All Other Values* and under *New Value* fill in value 0. Click *Add*, then *Continue* and finally *OK*. Follow the same steps to create the second dummy for group of diabetic subjects (remember to use 2 as the *Old Value*). Alternatively, use the *Transform* $\rightarrow$ *Compute variable* option to create dummies.