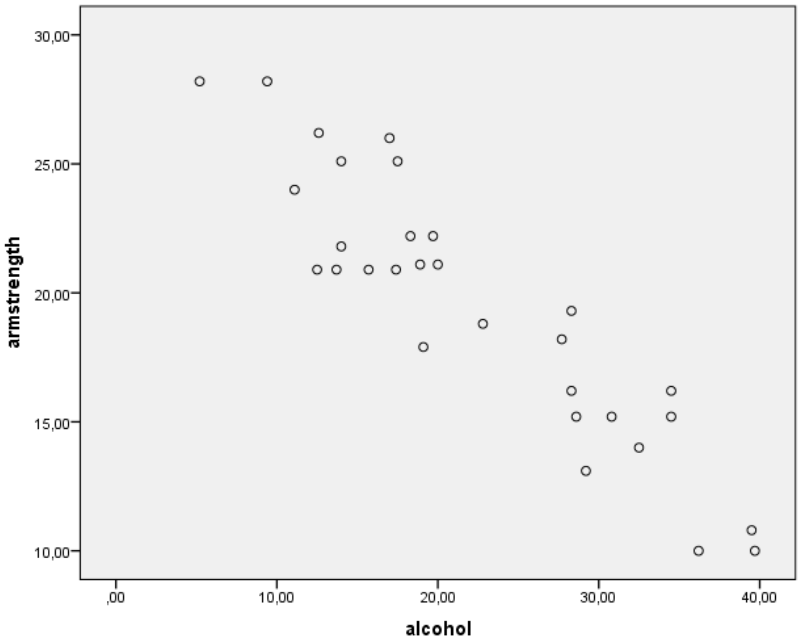
Practical "Correlation and simple linear regression"
P5

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* → *Legacy Dialogs* → *Scatter* and there select *Simple scatter*.

   Suggested answer:

```
GRAPH
  /SCATTERPLOT(BIVAR)=alcohol WITH armstrength
  /MISSING=LISTWISE.
```



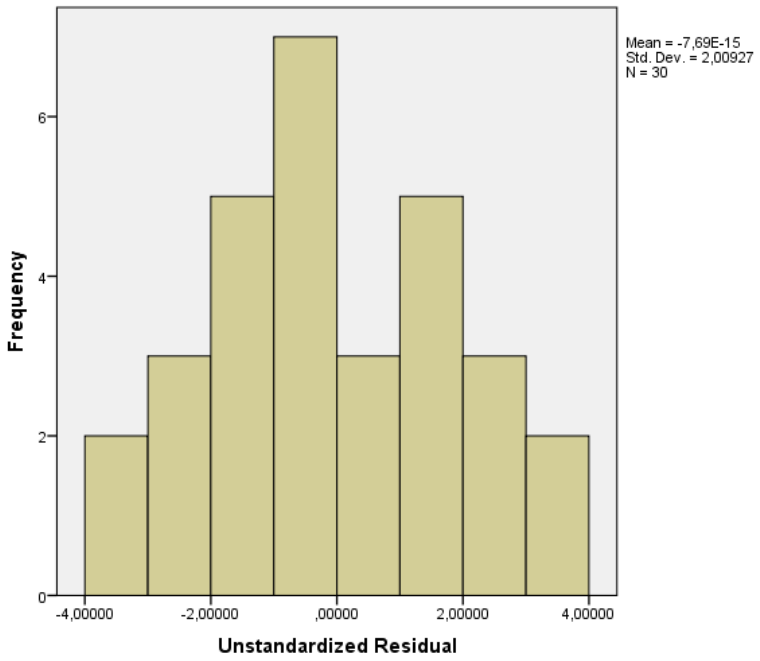
   The scatter plot shows that there is roughly a linear relationship between arm strength and alcohol intake, so a linear fit seems appropriate.

   (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* → *Regression* → *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* → *Descriptive Statistics* → *P–P Plots* and there select unstandardized residuals as a variable (no need to change other options). To plot the residuals against *alcohol*

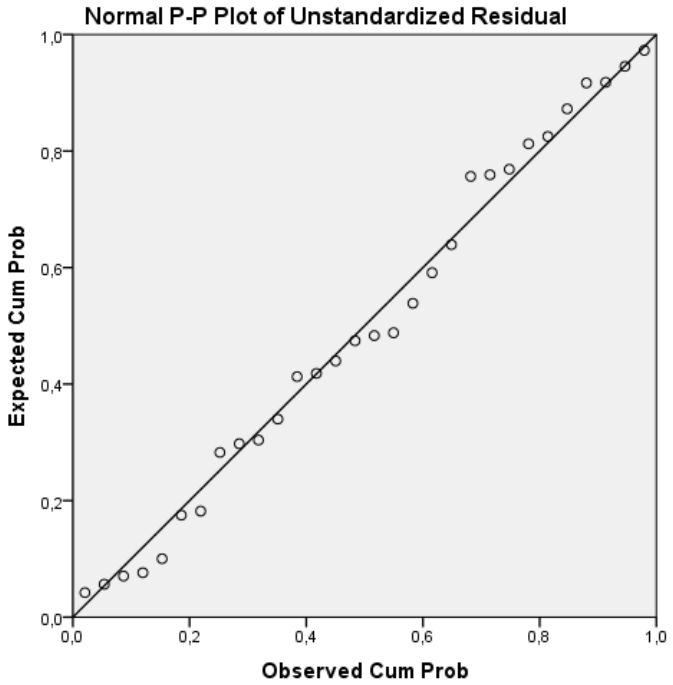
go to *Graphs* → *Legacy Dialogs* → *Scatter/Dot (Simple)* with the residuals as the y-axis variable and *alcohol* as the x-axis variable. In a similar way make a plot of the residuals versus the predicted arm strength values (PRE_1).

Suggested answer:

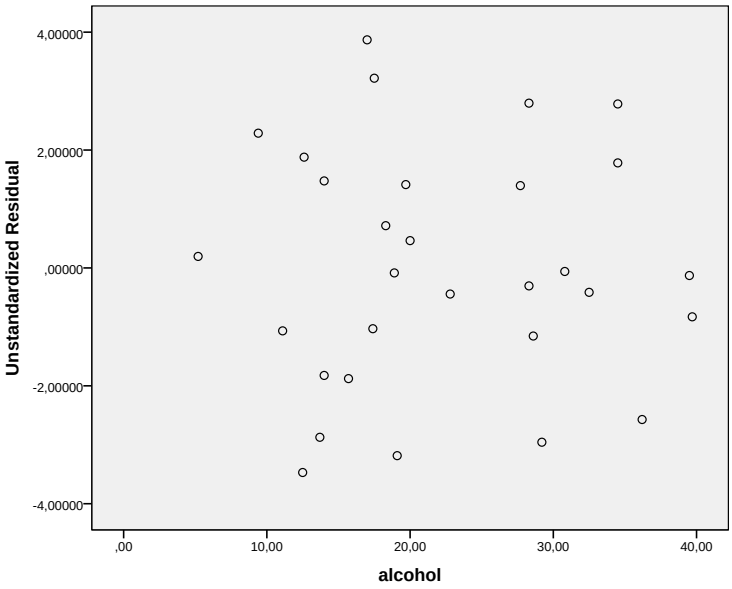
```
GRAPH
  /HISTOGRAM=RES_1.
```



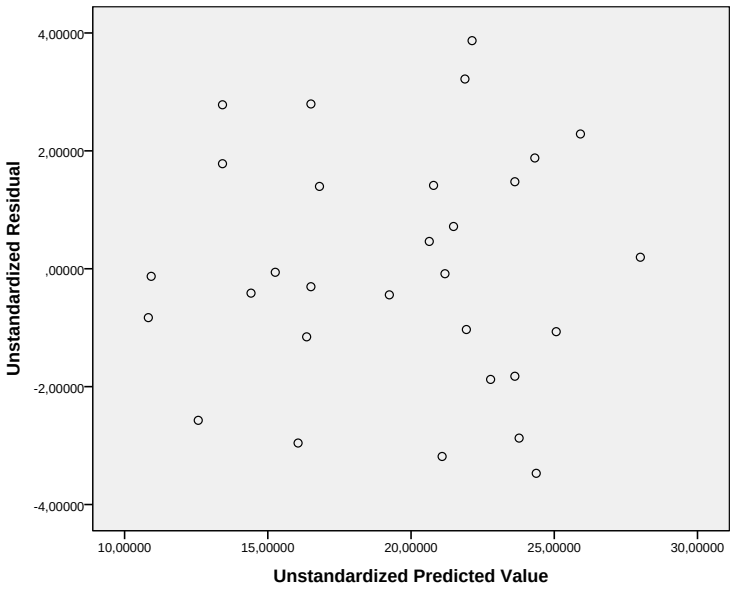

```
PPLOT
  /VARIABLES=RES_1
  /NOLOG
  /NOSTANDARDIZE
  /TYPE=P-P
  /FRACTION=BLOM
  /TIES=MEAN
  /DIST=NORMAL.
```




```
GRAPH
  /SCATTERPLOT(BIVAR)=alcohol WITH RES_1
  /MISSING=LISTWISE.
```




```
GRAPH
  /SCATTERPLOT(BIVAR)=PRE_1 WITH RES_1
  /MISSING=LISTWISE.
```



The observations are independent since there is only one pair of observations per patient. The histogram of residuals appears roughly normal (unimodal, fairly symmetric and without clear skewness or outliers). Further, the normal probability plot of the residuals shows the points close to a diagonal line. Therefore, there is no strong evidence against the normality assumption (though given the small sample size it is difficult to assess this). The plot of the residuals versus the explanatory variable shows that the average of the residuals remains approximately 0, the variation of the residuals appears to be roughly constant and there are no excessively outlying points, which supports the previous conclusions drawn from the scatter plot concerning the linearity of the association. The plot of the residuals versus the predicted values shows a random scatter of the points with a fairly constant spread and no excessively outlying points, so we will assume that the variance is constant. Thus, the assumptions for regression analysis appear to be met.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT armstrength
  /METHOD=ENTER alcohol
  /SAVE PRED RESID.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	30,594	,967		31,629	,000
	alcohol	-,498	,040	-,920	-12,436	,000

a. Dependent Variable: armstrength

Regression equation: *armstrength* = 30.594-0.498·*alcohol*. The slope of the regression line is –0.498. Thus, arm strength is expected to decrease by 0.498 kg (∼ 500 g) per each 1 kg/kg increase in alcohol intake.

(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?

Suggested answer:

At the α = 0.05 level of significance, there exists enough evidence to conclude that the slope of the population regression line is not zero (*p*–value $<$ 0.001). Thus, we conclude 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*.

Suggested answer:

The previous output includes the following table:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,920[a]	,847	,841	2,04484

a. Predictors: (Constant), alcohol

b. Dependent Variable: armstrength

We can additionally compute the Pearson correlation coefficient for *armstrength* and *alcohol*:

```
CORRELATIONS
  /VARIABLES=alcohol armstrength
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

		alcohol	armstrength
alcohol	Pearson Correlation	1	-,920**
	Sig. (2-tailed)		,000
	N	30	30
armstrength	Pearson Correlation	-,920**	1
	Sig. (2-tailed)	,000	
	N	30	30

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

The absolute value of the Pearson correlation coefficient corresponds to the value of *R* in the previous table, since there is only one explanatory variable in the regression model. In addition, we obtain results for a test on the correlation coefficient, concluding that the correlation between *armstrength* and *alcohol* is significantly different from zero (correlation coefficient –0.92, $p < 0.001$). From the value of the *R–square* we can say that approximately 85% of the variation in *armstrength* can be attributed

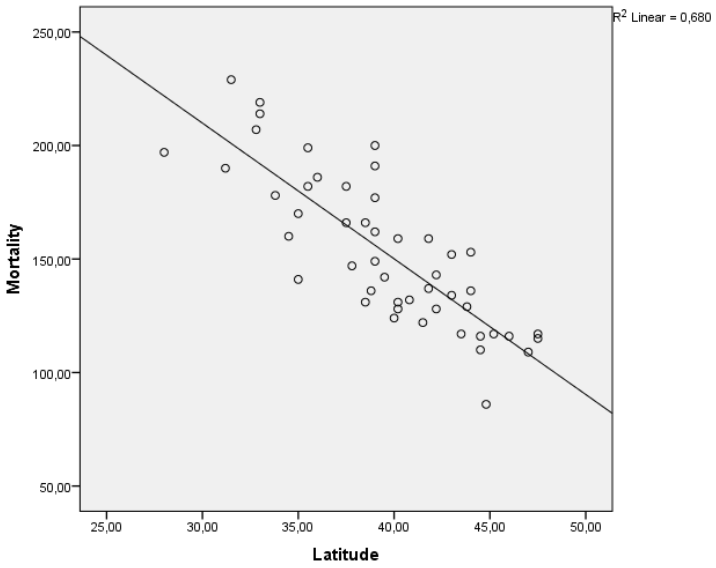
to its linear relationship with *alcohol*. The regression equation appears to be useful for making predictions since the value of *R–square* is close to 1.

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*.

   Suggested answer:

```
GRAPH
  /SCATTERPLOT(BIVAR)=Latitude WITH Mortality
  /MISSING=LISTWISE.
```




```
CORRELATIONS
  /VARIABLES=Mortality Latitude
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlations

		Mortality	Latitude
Mortality	Pearson Correlation	1	-,825**
	Sig. (2-tailed)		,000
	N	49	49
Latitude	Pearson Correlation	-,825**	1
	Sig. (2-tailed)	,000	
	N	49	49

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

The scatter plot suggests that there is a rough linear relationship between *mortality* and *latitude*. As the latitude increases the melanoma mortality rate tends to decrease. The correlation coefficient is –0.825, which indicates that there is a strong negative linear 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* → *Fit line at total* (change nothing) → Close the Chart Editor.

Suggested answer:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Mortality
  /METHOD=ENTER Latitude.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	389,189	23,812		16,344	,000
	Latitude	-5,978	,598	-,825	-9,990	,000

a. Dependent Variable: Mortality

The slope coefficient of the linear regression model of mortality onto latitude is –5.98. This means that for every additional one-degree increase in latitude the skin cancer mortality rate is expected to decrease by roughly 6 deaths per 10 million people.

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?

Suggested answer:

The expected change in the melanoma mortality rate associated with 5 degree increase in latitude is roughly –30 deaths (= $-5.978 * 5$) per 10 million people. Thus, the mortality rate in a state 5 degrees north of the other is expected to be lower by 30 cases. Equivalently, the mortality rate in the state 5 degrees south of the other is expected to be higher by 30 cases.

(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* → *Regression* → *Linear*. Click *Save* and select *Unstandardized* predicted values. Go back to the data and check the predicted mortality value in the last row.

Suggested answer:

For latitude of 40 degrees, the regression line predicts mortality rate of about 150 cases (= 389.198–5.978 ∗ 40) per 10 million people. This is a pretty reliable prediction for two reasons. First, the correlation between mortality and latitude is strong. Second, the latitude of 40 degrees is in the middle of the range of latitude values covered by the actual data (interpolation).

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).

   Suggested answer:

   We had obtained:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
logbtg	Equal variances assumed	,479	,496	-3,804	22	,001	-,95728	,25165	-1,47916	-,43540
	Equal variances not assumed			-3,804	21,900	,001	-,95728	,25165	-1,47930	-,43526

   (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*→ *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* → *Compute variable* option to create dummies.

Suggested answer:

```
RECODE group (1=1) (ELSE=0) INTO norm.
EXECUTE.
RECODE group (2=1) (ELSE=0) INTO diab.
EXECUTE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT logbtg
  /METHOD=ENTER norm diab.
```

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5,498	1	5,498	14,471	,001[b]
	Residual	8,359	22	,380		
	Total	13,857	23			

a. Dependent Variable: logbtg

b. Predictors: (Constant), diab

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,433	,178		13,675	,000
	diab	,957	,252	,630	3,804	,001

a. Dependent Variable: logbtg

The mean difference in (log transformed) urinary beta-thromboglobulin excretion between groups (=0.957) is the same. Also, the value of the t-statistic (=3.804) is the same (and its degrees of freedom (=22) and the associated p-value (=0.001)). Notice that the estimates of the aforementioned statistics differ in sign in the two analyses, which is caused by using a different reference category for the *group* variable in the linear regression and independent samples t-test. After resolving issue of the reference category, we can conclude that the independent samples t-test is equivalent to linear regression with one dummy for the grouping variable.
NOTE: In this particular case, in the corresponding ANOVA table that is displayed when running a linear regression, we see that the F–statistic of 14.471 is the square of the t-statistic of 3.804. This is because a squared Student–t distribution with df degrees of freedom follows an F–distribution with 1 and df degrees of freedom.