Lab 8: ANOVA, Correlation & Regression

Lab Objectives:

- Compare the means of multiple groups using Analysis of Variance (ANOVA)
- Make post-hoc comparisons with the Tukey-Kramer test
- Analyze the association between two numerical variables using correlation and regression

Exercise 1: ANOVA

A **One Way ANOVA** is an analysis of variance in which there is only one independent variable. It is used to compare mean differences in two or more groups. An ANOVA with two groups is mathematically equivalent to a two-sample two-tailed t-test. A significant ANOVA result implies that at least one group has a different mean from one other group, but does not directly tell you what the significant difference is. Post-hoc comparisons, such as the **Tukey-Kramer test**, can be used to determine which group or groups are significantly different from the others.

ANOVA assumes equal variance of each group, that each group has a normal distribution of the variable, and that each population is randomly sampled. The **Kruskal-Wallis test** is a nonparametric test that compares multiple groups. It is often used in place of ANOVA when the assumptions of ANOVA are violated.

Maternal Role Adaptation

Various factors can influence how well new mothers adapt to a maternal role. Low birth weight of a baby due to premature labour is believed to increase psychological stress and risk of depression in new mothers, reducing maternal role adaptation. In this study, maternal role adaptation was compared in a group of mothers of low birth-weight (LBW) infants, mothers of LBW infants who had an experimental intervention, and mothers of full-term infants. The hypothesis was that mothers of LBW infants in the experimental intervention would adapt to their maternal role as well as mothers of healthy full-term infants, and that each of these groups would adapt better than mothers of LBW infants in the control group. A lower maternal role adaptation score indicates better adaptation.

- Open the file *maternal role adaptation.sav*

- ANOVA is robust to deviations from normality, particularly when sample sizes are large. It is also robust to departures from the assumption of equal variance, but only if the samples are all large, about the same size, and there is no more than about a tenfold difference among variances. However, it is still a good idea to check whether your data meet these assumptions. Begin by analyzing the data using the descriptive statistics, explore command for deviations from normality.

- Select Analyze, Compare Means, **One-Way ANOVA**.

- Add the “adapt” variable to the dependent list, and the “group” variable to the factor list. This will tell the test that we are interested in comparing the adaptation scores between the different groups.

- Click "**Post Hoc**" and put a check-mark next to **Tukey**.
- Click "**Options**" and put check-marks next to Descriptive, Homogeneity of variance test, and Means plot.
- Click OK to run the test.
- The first table shows Descriptive statistics for your samples, including N, mean, standard deviation, standard error, and 95% confidence interval for the mean.
- The second table contains the results of a Leven's test for equal variances. The null hypothesis of this test is that the samples have equal variances.
- The third table is the ANOVA table. We will learn how to calculate the important values in an ANOVA table in class. The table shows you the calculated variance ratio (F) and the corresponding P-value under "Sig". A P-value less than 0.05 indicates that mean adaptation score (in this case) differs among treatments. However, this alone does not tell us which group or groups are significantly different.
- To determine how the groups differ, view the next two tables which contain the results of the Tukey-Kramer post hoc test, which compares all pairs of means. The first table is the Multiple Comparisons table. It shows the mean difference for each combination of groups, the standard error of the difference, and the significance value (P-value). Note that each combination of groups appears in the table twice. These are highlighted in the same colour in the figure below. For example, the two comparison highlighted in blue are the same (LBW experimental compared to LBW control). SPSS makes things easier by indicating comparisons that are significant with an asterisk (*).

Dependent Variable: Adapt

Tukey HSD

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
LBW-Exp	LBW-Control	-6.593*	.735	.000	-8.34	-4.84
	Full-Term	-1.162	.681	.209	-2.79	.46
LBW-Control	LBW-Exp	6.593*	.735	.000	4.84	8.34
	Full-Term	5.430*	.695	.000	3.77	7.09
Full-Term	LBW-Exp	1.162	.681	.209	-.46	2.79
	LBW-Control	-5.430*	.695	.000	-7.09	-3.77

*. The mean difference is significant at the 0.05 level.

- The next table under Homogeneous Subsets called "Adapt" is also useful. The results of the Tukey-Kramer test are exact when the sample size in every group is the same. If sample sizes are different, then the test is conservative, which means that the probability of making at least one Type I error is lower than the stated alpha. Groups that fall into the same subset are not significantly different from one another (see below). The values shown under each subset is the mean for that group. So in this case, the means for the Full term and LBW experimental groups

are not significantly different from each other. However, both are significantly different from LBW control, which falls into subset 2 by itself.

Adapt

Tukey HSD[a,b]

Group	N	Subset for alpha = 0.05	
		1	2
LBW-Exp	29	13.00	
Full-Term	37	14.16	
LBW-Control	27		19.59
Sig.		.230	1.000

Means for groups in homogeneous subsets are displayed.

- The means plot allows you to visualize the difference between group means using a graphical method. If you examine this plot you can quickly see that the LBW Control group had a higher adaptability score (indicating poorer adaption to a maternal role) than the LBW experimental and Full term groups.

➢ Answer questions 1 and 2 on Sakai.

Circadian clock

In a 1998 study, Campbell and Murphy reported that the human circadian clock can be reset by exposing the back of the knee to light, which was met with much skepticism. A later experiment re-examined the phenomenon using 22 people who were awakened from sleep to receive a 3-hour light treatment of the eyes, knees, or neither (control). Effect on circadian rhythm was determined by measuring melatonin production two days later. Melatonin secretion is tightly connected to circadian rhythm, with secretion beginning to rise about two hours before a person's regular bedtime. A negative measurement indicates a delay in melatonin production, while a positive number indicates an advance. The data is found in the file *Knees circadian clock.sav*. Use a one-way ANOVA to assess whether light-treatment affected circadian rhythm.

➢ Answer questions 3 and 4 on Sakai.

Memory recall

In a study of memory recall, a group of subjects were asked to read through a list of 27 words. One group was assigned to count the number of letters in each word. Another group was assigned to think of a word that rhymed with each word. The third group was instructed to think of an adjective that could be used to modify each word. An imagery group was asked to form vivid images of each word. These first four groups were not told that they would later have to recall the word list. Finally, the last group ("Intentional") was asked to memorize the words as best as they could for later recall. Each group reviewed the list of 27 words three times, and was then asked to write down as many words as they could remember. The data is found in the file *Recall.sav*. Examine the data to determine if there is any difference between the groups.

➢ Answer questions 5 – 8 on Sakai.

Cuckoos Eggs

The European cuckoo does not look after its own eggs, but instead lays them in the nests of birds of other species. Do cuckoos lay eggs of different sizes in nests of different hosts? The data file *cuckoo eggs.sav* contains data on the lengths of cuckoo eggs laid in a variety of other species' nests. Examine the data to determine if there is any difference between the groups. Use a Tukey-Kramer test to determine which pairs of host species are significantly different from each other.

- Answer question 9 on Sakai.

Exercise 2: Kruskal-Wallis test

If the data do not meet the assumptions of the ANOVA test (normal distribution and equal variances in all populations), and the data are not improved by data transformation, a **Kruskal-Wallis test** can be used instead.

Malaria & Maize

The pollen of the corn (maize) plant is known to be a source of food for larval mosquitoes of the species *Anopheles arabiensis,* the main vector of malaria in Ethiopia. The production of maize has increased substantially in certain areas of Ethiopia recently, and over the same time malaria has entered in to new areas where it was previously rare. This raises the question: Is the increase of maize cultivation partly responsible for the increase in malaria?

Data in the file *malaria maize.sav* contains information on the level of cultivation of maize (low, medium or high) and the rate of malaria per 10,000 people for several sites in Ethiopia. Assume that you know from prior analysis that the incidence rate of malaria is not normally distributed.

- Open the file *malaria maize.sav*
- Since we know the data are not normally distributed, and our samples are quite small, we will conduct a non-parametric test. Select Analyze, Nonparametric tests, Independent samples.
- On the Objective tab, select “customize analysis”.
- On the Fields tab, add Incidence rate to the test fields, and Maize yield to the groups.
- On the Settings tab, select “customize test”. Then check the option next to Kruskal-Wallis 1-way ANOVA. In the multiple comparisons dropdown menu, ensure “All pairwise” is selected. Click Run.
- In the output window you will see a hypothesis test summary box. This box makes it easy to see the null hypothesis, the significance value, and the decision. However, it does not show us the pairwise comparisons to see which groups are different from one another.
- Double click on the hypothesis box to open a new window called model viewer.
- In the window that opens, you should be able to see box plots representing the data.

- In the right-hand window at the bottom next to "View", select "Pairwise comparisons" from the dropdown.

- The resulting table shows us the significance value for each pairwise comparison, and an adjusted significance value according to the number of pairwise tests made. This adjusted significance helps keep the probability of type I error (α) at 0.05 total across all of the comparisons. Note that in this case, since we are conducting three pairwise comparisons, the adjusted significance is the calculated significance x 3.

➢ Answer question 10 on Sakai.

Nematode lifespan

An experiment in 2005 examined the effect of the anticonvulsant medication trimethadione on the lifespan of the nematode worm C. elegans. The study compared the effect of treatment provided at the larval stage, the adult stage, and both stages, to control worms treated with water only. Examine the data found in the file *nematode.sav* for deviations from the assumptions of a one-way ANOVA test. If the data violate these assumptions, perform a Kruskal-Wallis test to determine whether any of the treatments had a significantly different lifespan compared with the water control.

➢ Answer question 11 on Sakai.

Exercise 3: Correlation

When two numerical variables are associated, we say they are correlated. The **correlation coefficient** is a quantity that describes the strength and direction of an association. It reflects the amount of "scatter" in a scatter plot of the two variables, but unlike linear regression does not measure how steeply one variable changes when the other changes. The maximum correlation possible is 1.0 or -1.0. A correlation of 1.0 indicates that the measurements lie along a straight linear line and show a positive relationship (as one variable increases, the other increases as well). In comparison, a correlation of 0.5 will have a generally positive relationship, but a lot more scatter among the points. A correlation of -1.0 indicates that the measurements lie along a straight linear line and show a negative relationship (as one variable increases, the other decreases).

Correlation analysis assumes that the measurements have a bivariate normal distribution, which is a normal distribution in two dimensions rather than one (see Figures below). A bivariate normal distribution occurs when the relationship between X and Y is linear, the cloud of points on a scatter plot has a circular or elliptical shape, and the frequency distributions of X and Y separately are normal. Visualizing a scatter plot of the data is an important step to detect deviations from these assumptions. Histograms depicting the frequency distributions of X and Y separately are also helpful.

If the assumptions of correlation analysis are violated, then **data transformation** may be tried. If data transformation does not improve the fit of the data to bivariate normality, then **Spearman's rank correlation** is used instead.

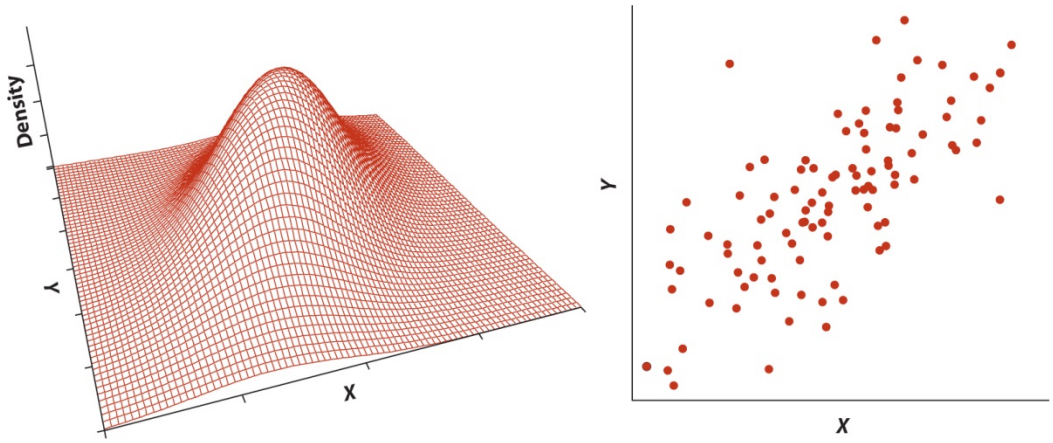


Figure 16.3-1. A bivariate normal distribution.

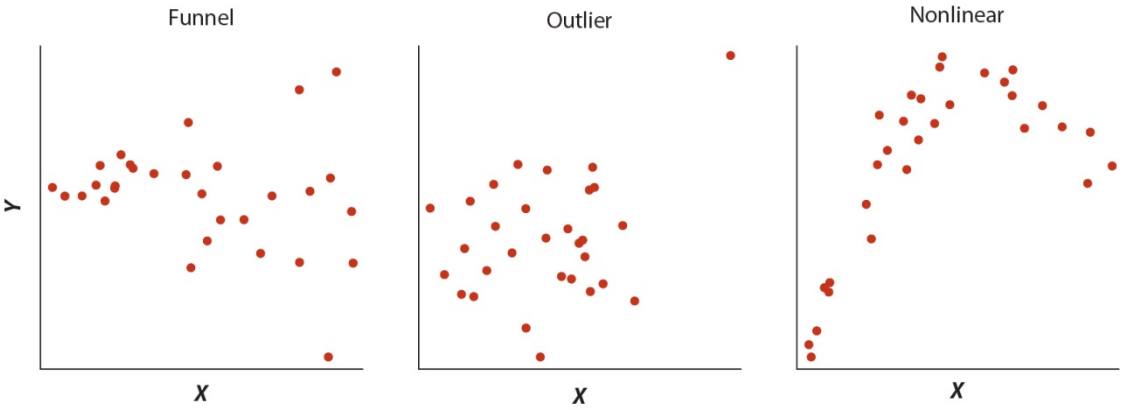


Figure 16.3-2. Examples of data that deviate from a bivariate normal distribution.

Height & weight by gender

- Data comparing the height and weight of a sample of males and females is found in the file *Heightweight.sav.* Open this file now.
- In this case, we know from prior knowledge that the variables height and weight are normally distributed. Construct a scatter plot to determine whether the relationship between the two variables appears to fit a bivariate normal distribution.
- Select Analyze, Correlate, Bivariate. This option will examine the data for a simple association between two variables.
- In the dialog box, add height and weight to the variables list. Under correlation coefficients, ensure "**Pearson**" is checked. Select "Two-tailed" for Test of significance, and check the box next to "flag significant correlations". **Note** that if you are working with data that violate the assumptions of correlation analysis, you can conduct a Spearman's rank correlation by checking the box next to "**Spearman**" instead of Pearson.

- Click on options, and check the box for "Means and standard deviations". Click continue, then OK to run the test.
- In the output window, you will first see a summary of the descriptive statistics, followed by the correlation table. The Pearson Correlation coefficient is a measure of the direction and strength of the relationship between the two variables. The significance value indicates whether or not this association is statistically significant.
- The data contains data for both males and females. If we want to conduct subgroup correlations for males and females separately, the easiest way to do this is to split our data file. Select data, split file.
- Select organize output by groups.
- Add sex to the box "groups based on". Click OK.
- Select analyze, correlation, bivariate again. The same variables and options you used last time should still be selected. Click OK to analyze the data.
- The descriptive statistics and correlations should now show up separately for males and females.
- To construct a scatterplot of the data, we will first un-split the file. Select data, split file, and then highlight the option "analyze all cases, do not create groups". Click OK.
- Select graphs, chart builder, Scatter/dot. Select the first option "simple scatter". Add height to the X-axis and weight to the Y-axis. Click OK.
- The resulting chart allows you to visualize the scatter of data. Double click to open the chart and use the correct button in the menu to add a linear fit line.
- Adding a linear fit line causes an R^2 statistic to show up in the margin. This statistic is used in regression analysis. It measures the fraction of the variation in Y that is explained by X. The two values are closely related – the correlation coefficient is R, and R-Square is simply this value squared

➢ Answer question 12 on Sakai.

Chocolate & Nobel Prizes

There is evidence that higher consumption of foods containing chemicals called flavonols (including cocoa, red wine, green tea, and some fruits) increases brain function. Messerli (2012) examined whether chocolate consumption in a country is correlated with the number of Nobel prizes. The data are found in *chocolate.sav*. Begin by constructing a scatter plot to view the data. You should see that the cloud of points are funnel shaped (wider at one end then the other), which indicates deviation from bivariate normality. Imagine that you try many types of data transformation that does not improve the fit of the data to bivariate normality. Conduct an appropriate correlation test to analyze the association between the two variables.

➢ Answer question 13 on Sakai.

ADD, IQ, and GPA

Open the file *appendixD.sav*, which contains data that you have previously analyzed this term. Calculate the correlation coefficients for ADD score, IQ, GPA and English grade. Note that you can add all variables at the same time, and SPSS will compute the correlation coefficient and significance for each pair of variables. Assume that there are no deviations from bivariate normality in this case.

- Answer question 14 on Sakai.

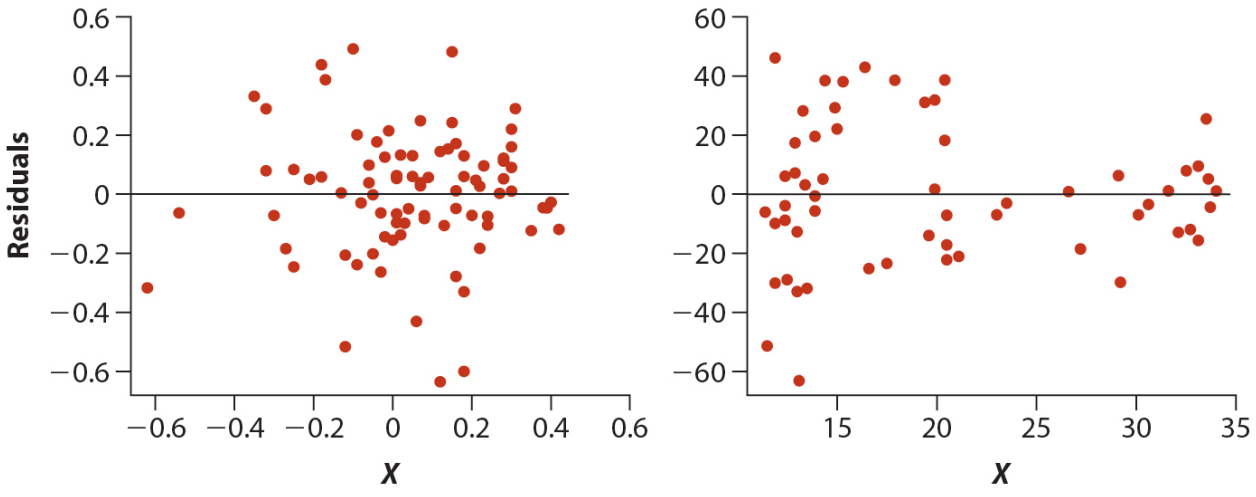
Exercise 4: Regression

Linear regression uses a line to predict a response numerical variable (Y) from a numerical explanatory variable (X). The regression line will take the form Y = a + b X. In this equation, a is the intercept (where the line crosses the axis at X = 0), and b is the slope of this regression line.

A common statistic used in regression analysis is R^2. R^2 measures the fraction of variance in Y that is predicted by X. If R^2 is close to one (the maximum possible value), then X predicts most of the variation in the values of Y, and the Y observations are tightly clustered around the regression line with little scatter.

Hypothesis tests about regression slopes can be made using either a t-test or ANOVA approach. A P-value of less than 0.05 indicates that the slope between the two variables is significantly different from zero.

Linear regression assumes that the true relationship between X and Y is linear, that for every value of X the corresponding values of Y are normally distributed, and that the variance of Y values is the same for all values of X. Non-linearity is best detected by visualizing a scatter plot of the data. Non-linear relationships between X and Y can often be examined by the use of transformations. Non-normality and unequal variance is best examined with a residuals plot, which calculates the difference between every Y data point and the predicted Y according to the regression equation, and plots this against X. A residuals plot should have a roughly symmetric cloud of points above and below zero, a higher density of points close to zero, and no noticeable change from left to right. The plot on the **left** below fits these requirements well, while the plot on the right does not.



The Lion Nose

- Whitman et al (2004) examined the relationship between age and the proportion of black pigmentation on the noses of 32 male lions of known age in Tanzania. The goal was to be able to predict the age of a lion from the amount of black on its nosepad. Open the file *Lions ages.sav* to view the data.

- It is always good practice to examine your data graphically before conducting statistical analysis. Construct a scatter plot to view the relationship between proportion black (X axis) and lion age (Y axis). In the chart editor, add a linear fit line. This approach quickly allows you to view the linear equation describing the relationship between two variables, as well as the R^2 value.

- Select **analyze**, **regression**, **linear**.

- Add proportion black to the independent variable, and age to the dependent variable. The variables are added in this order because we are interested in predicting a lion's age from the proportion of black pigmentation on its nose, which can be measured.

- Under **statistics**, check the boxes for estimates, confidence intervals, model fit, and descriptives. Click continue.

- Under **save**, check the boxes for predicted values unstandardized, and residuals unstandardized. Click continue, then click OK to run the linear regression.

- In the output window, you should see a box with descriptive statistics, followed by a box with the correlation data. This is similar to the correlation output calculated previously through analyze, correlate, bivariate. However, note that the significance is given for a one-tailed test, rather than a two-tailed test.

- The next two tables contain information on the variables used in the linear regression model. In the model summary table, you can see the R value (correlation coefficient). It should be the same as the value in the correlation table above. You can also see the R^2 value, which should be the same as previously viewed in your scatter plot. Note that both positive and negative correlations will produce a positive R^2 value, which allows them to be compared on the same scale. We will not use adjusted R^2values, so you can ignore this box.

- The ANOVA table contains the results of a hypothesis test using the ANOVA approach. The null hypothesis being tested is that the slope between the two variables is zero (no relationship). The P-value of this test is found under "Sig.". If this value is less than 0.05 then there is evidence for a linear relationship between the two numerical variables.

- The final table is the Coefficients table, which is what we are most interested in. In the first row labelled constant, the value under B is the estimated constant value in the equation Y = a + b X (the intercept). The standard error and confidence interval for this value are given. Do not worry about the t-statistic and significance in this row, as we are not interested in the intercept, but rather the slope.

- The next row labelled "proportion.black" contains the estimated value for the slope in the B column. The standard error and 95% confidence interval of the slope are also given.

- The proportion.black row also contains the results of a hypothesis test using the t-test approach. The t-statistic for the slope is calculated from the formula: $t = \frac{b-\beta_0}{SE_b}$ where b is the estimate of the slope in the sample, β_0 is the null hypothesized value of the slope (usually zero) and SE_b is the standard error of the slope in the sample. Since we are only working with two variables, the P-values from both the ANOVA approach and t-test approach should be identical.

- Now return to the data view in your file. You should see that the regression analysis added two columns to your data. The first column, labeled "PRE_1", provides predicted ages according to the regression formula and the proportion of black observed on the nose. For example, the first lion has an age of 1.1, and a proportion black of 0.21. According to the regression formula, a lion with 0.21 black on his nose would be predicted to be 3.11 years old.

- The second column, labeled "RES_1", provides the residual for each data point. Residuals are calculated by subtracting the predicted value of Y from the actual value of Y, for each data point. For example, for the first lion, the predicted age is 3.11 and the actual age is 1.1. Therefore the residual is -2.01.

- The residuals data can be used to construct a residuals plot, which lets us detect non-normality and unequal variance. Construct this graph now by plotting proportion black (X axis) against residuals (Y axis). You can add a horizontal line at zero, using the linear fit line function. In this case there are fewer points to the right of the graph (high proportion black), which may make the spread of points look different from left to right, but it is not significant enough for us to conclude that the distribution of Y values at each value of X is not normal or has unequal variance.

- Another way to evaluate the assumption that the residuals are normally distributed is to analyze the residuals data with a normality test and histogram. Try this now.

➢ Answer question 15 and 16 on Sakai.

Brain size

Some nonlinear relationships can be made linear with a suitable transformation. One of the most common transformations is the log transformation. The data in the file *mammals.sav* contains species name, body mass (in kg) and brain size (in g) of 62 different mammal species. Open this data and construct a scatter plot to visualize brain size (on the Y-axis) compared to body mass. You should recognize immediately that this scatter of points does not look like a normal linear relationship. Transform each variable using the natural log transformation, as you learned last week. Re-plot the data using the log transformed variables. If this improved the linear relationship between the two variables conduct a linear regression analysis on the log-transformed variables.

➢ Answer question 17 on Sakai.

Telomeres

The ends of chromosomes are called telomeres. As individuals age, their telomeres shorten, and there is evidence that shortened telomeres may play a role in aging. Telomeres can be lengthened in germ cells and stem cells by an enzyme called telomerase, but this is not expressed in most healthy somatic cells. (Cancer cells, on the other hand, usually express telomerase.) A set of data collected by Nordfjäll *et al.* (2005) examined whether there is a relationship between telomere length of fathers and their children. Examine the data in *telomeres.sav* by regression to determine whether offspring telomere length can be predicted from father telomere length.

- Answer question 18 and 19 on Sakai.

Stress and mental health

Open the file *symptoms and stress.sav.* This file contains information on stress levels and severity of mental health symptoms in 106 patients. The authors are interested in whether the severity of mental health symptoms can be predicted from stress levels.

- Answer question 20 on Sakai.

Further reading: Multiple Regression

Multiple regression can analyze data for a linear relationship between several independent variables and one dependent variable. For example, you may be interested in predicting the height of plants (dependent) based on soil water content, soil nitrogen content, and sunlight level (independent variables). If you are working with multiple numerical explanatory variables for your research project, there is a great tutorial on conducting multiple regression in SPSS that can be found at:

http://www.ucdenver.edu/academics/colleges/nursing/Documents/PDF/MultipleRegressionHowTo.pdf

Click cancel on the log in box if it pops up, the link should still load!