Chapter 6

6.1

While linear regressions are most commonly used, many relationships are simply non-linear. In these situations a linear regression analysis would show that there is no relation between the variables, while in fact there is simply a non-linear relationship. This makes it important to understand non-linear regression as well.

First of all we can examine whether a relationship between variables is likely to be linear or not from examining the scatterplot. Next we need to fit the model. Obviously non-linear regression has need of a different type of equation. Specifically, the normal regression equation (of a straight line; y = b1x1 + c) needs to be transformed to become curvilinear. This is done by squaring the independent variable (x ◊ x2; accordingly the values of the independent variable are all squared as well: 4 ◊ 16, 12 ◊ 144, etc.). Thus the equation we get is: y = b1x1 + b2x12 + c.

Squaring the independent variable (x2) leads to a quadratic function, which is curvilinear but only contains one curve. For some models it's needed to have a line that contains multiple curves. A function where the line has two curves is called a cubic function, and goes as follows: y = b1x1 + b2x12 + b3x13 + c.

Keep in mind that in all three of these models there is only one independent variable, it is simply brought up in some of the equations multiple times, with different b-coefficients.

To show a line with an upward exponentially decreasing curve, you need to use a logarithmic function. A logarithm (log) is usually put to base 10. This means that the log10 of any number is the number that when placed in the power of 10 creates the original number (say log10(a) = b, because a = 10b). Because such a function has the effect of making the differences between higher values smaller than the differences between lower values, it creates a curve that gradually levels of its slope while going upwards.

Finally one can make use of an inverse function, which is calculated as follows: y = 1/x. The inverse function makes small values large, and large values small.

Points to keep in mind while carrying out a non-linear regression analysis:

- You can add a constant, as long as it is added to every value of the variable, so it doesn't affect the interpretation of the results. The same goes for multiplying the variables. As long as such changes are made consistently the standardised estimates, probability values, and thus conclusions, do not change.
- Transformations affect the shape of the function, but also the distribution of the variable. Meaning that a transformation could cause a variable to not be normally distributed anymore, or could even cause outliers to appear. Any solution here involves a trade-off or a compromise, and usually involves deleting problematic cases and/or values from the analysis.
- The values of the independent variable itself also effect the curve of the function.

Not all transformations are possible. A log of the number 0, for example, yields no value. Therefore, you sometimes need to add a constant in order for the range of numbers to function as they should.

If a linear model and a non-linear model both describe the data with the same accuracy, go with the linear model as it is simpler.

Never use stepwise regression when it comes to non-linear variables in a model, as such techniques favour non-linear variables over linear ones, and thus end up making the model more complex than necessary.

6.2

Regression analysis works best on a dependent variable that has a continuous string of values. But there are also plenty of situations where a categorical dependent variable (data consisting of only yes/no for example) may need to be examined. To do this standard regression techniques need to be adapted. After all, if only two situations are possible, as defined by the dependent variable, then most standard outcomes of a regression analysis, like slopes and distributions, make no sense.

To analyse such data the data itself needs to be transformed. This differs from the earlier transformations discussed as here we are applying this transformation not to the independent variable, but to the dependent variable. The transformation needed here is a logit transformation, which means that the regression analysis based on this transformation is called a logistic regression.

One way of transforming the data in this manner is by using probabilities instead of the categorical variable. For example, let's say we're analysing several groups with the dependent variable being whether they have pets. The categorical variable would show a 1 for having a pet, and a 0 for not having a pet, giving each group an absolute number of the amount of people in the group with pets. Using probabilities would mean giving each group a number between 0 and 1 signifying the chance that someone in that group has pets (by dividing the absolute number by the number of people in the group). In this manner there is a continuum.

It is only a limited continuum, however, as the numbers cannot be lower than 0 or higher than 1. Regression analysis would show number beyond this range, which are then useless. So while probabilities seem to offer a solution, we really want to turn to odds. Specifically we make use of the odds ratio, which show the odds of an event happening. It is calculated with the following formula:

$$Odds\ ratio = \frac{P(event)}{1-P(event)}$$

An odds ratio can be transformed into a probability in the following manner:

$$P(event) = \frac{odds(event)}{[1+odds(event)]}$$

While using odds means that no value predicted by a regression can be too high, it doesn't solve the problem of values beneath zero not being possible. Thus, one final stage is needed to make the transformation complete and to make regression of this data possible: calculating the logit. The logit is actually just the natural logarithm (log) of the odds ratio. Using the logit as the values for the variable makes it possible for the values to extend below zero.

Interestingly these three forms of values remain connected, meaning that with one of these, the other two can always be calculated:

Probability ↔ Odds ↔ Logit

Linear regression makes use of ordinary least squares (OLS) regression. This is regression that fits the best line to the data by squaring the residuals (discrepancy between data points and the regression line; the residuals are squared so get rid of negative numbers) and summing these up. The line where the resulting number is the lowest, is the line that best fits the data.

In logistic regression, however, OLS estimation cannot be used due to the type of equations that are present in this form of regression. An estimator called maximum likelihood (ML) is used here instead. Because of this use of ML logistic regression analysis creates an output different than that of an OLS analysis, but it is still similar in most ways.

ML estimation makes use of the log likelihood function. This is a measure that is calculated to indicate how well the parameter estimates (which are estimated by guessing) fit the actual data. According to this outcome the parameter estimates are adjusted and a new estimation is made of the log likelihood function. This tweaking continues until no further improvements can be made, the parameter estimates are kept.

It can be generally put that the larger the log likelihood function is, the better the model will fit the data.

As discussed before, a hierarchical assessment of variables in a regression is possible. Earlier it was addressed how to do this in an OLS regression, but it is of course also possible in multiple logistic regression. Luckily SPSS automatically carries out calculations to assess improvement to the model when a variable is added.