Chapter 2

2.1

We will look at the basic model of multiple regression. The model is all about:

DATA = MODEL + ERROR

Mostly when people make causal attributions about events, they tend to attribute an effect to one single cause. However, mostly there are several events that may have contributed to the result. The fact that people tend to think of one factor as being the cause of an event we refer to as the myth of monocausality. So we need to make a more complex mode. If we have two (or more) independent variables, we need to use this equation:

$$\hat{Y} = b_1 x_1 + b_2 x_2 + c$$

In this formula b1 is the slope for the first independent variable, and b2 is the slope for the second independent variable.

2.2

Sometimes we use the term ‘controller for’, but what does this mean? In experimental psychology, variables that we are not interested in are controlled using standardised conditions. But sometimes it is hard to control for everything, in these cases we use statistical control.

When we have one independent variable, we explain a certain proportion of the variance in the dependent variable. When we introduce a second independent variable, we want to know how much variance that variable explains. However there is a problem, the two independent variables are likely to correlate, and if they do, they will share some of their variances. The multiple regression calculates these proportions by taking into account the correlations between independent variables, and assessing the effect of each independent variable when the other variable have been removed.

2.3

So if we enter both the independent variables in the regression equation, we get estimates of the slope coefficients for each variable, controlling for the other variables.

2.4

R is the multiple correlations, sometime known as the coefficient of determination. It represents the total correlation between all the independent variables and the dependent variable. R2 is the value of R having been squared. The R2 represents the total amount of variance accounted for in the dependent variable by the independent variable(s). The value of R squared can be interpreted as the proportion of variance explained by moving the decimal point two places to the right and expressing this value as a percentage.

2.5

Adjusted R squared is a reduced value for R squared which attempts to make an estimate of the value of R2 in the population. The reason behind the adjustment is that if another independent variable is added it is very unlikely that the correlation between that independent variable and the dependent variable will be exactly zero, even if it is zero in the population. R2 will always go up a little when another independent variable is added. So adjusted R2 is adjusted down to compensate for this increase in R2. The smaller the sample size, the greater the random variation from zero will be and therefore the larger the downward adjustment in R2 is required. The calculation of adjusted R2 is as follows:

$$\text{Adj. } R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1} \qquad (102)$$

Where n is the number of people; k is the number of independent variables. As n increases, the amount by which R2 is adjusted downwards decreases, and as k increases, the amount by which R2 is a reduced increase.

2.6

Interestingly enough regression can also lead to an ANOVA output. An ANOVA is typically used to examine variability and is thus well equipped to determine the variance in the dependent variable, and the amount of this variance that is the consequence of the independent variables.

The value ‘F’ in such an output id the probability of getting the given R2 value if the actual value in the population is 0 (or: F is the probability of R2).

2.7

The constant is the expected value of the dependent variable when the independent variable (or: all independent variables) is/are zero.

Unstandardized coefficients subsequently tell us the increase in the dependent variable with every increase in the independent variable.

2.8

While the contribution of independent variables is usually assessed with all variables being in the same time frame, in some cases there is more certainty about cause-and-effect and/or hierarchy. Such hierarchy of variables can also be entered in a multiple regression analysis. (Note: This example refers to a situation where there is no designated dependent variable, just various independent variables). To do this separate regression analyses need to be done, varying the independent variables used, in order to determine whether two have a relationship that is statistically significant from the other relationships between the variables.

2.9

A parsimonious model is a model that explains the largest amount of variance in the dependent variable with the smallest amount of independent variables. The general aim of every statistical technique is to create such a model. Statistical computer programmes have added new computational abilities to such techniques. In this manner regression models can be built in a series of multiple steps, adding and/or removing independent variables. The techniques for this are as follows:

The backward technique. Here you start off with a model that includes all independent variables, and independent variables that do not meet the required parameters are removed.

The forward technique. Here you start without any independent variables in your model, and sequentially add the variable with the highest standardised beta-score. This is done until there are no variables with significant values anymore.

Stepwise regression, a combination of the backwards and forwards technique, using the manners outlined above to add and remove variables, and assess the model.

The regression that such techniques use is stepwise, however, which has the problem of creating high-biased R2-values. Adjusting the R2-value itself will not fix this, as the number of potential variables in the model are not taken into account. Besides this there is the fact that p-values are continuously altered, as the p-value is dependent on the amount of variables. This means that the significance of both the R2 and the individual beta values are incorrect. One final problem that has been identified with this method is that it created a large amount of paperwork.