Chapter 11: Multivariate Analysis

Progress Questions

1. A ***correlation matrix*** is a useful way of undertaking an initial exploration of the relationships between the important variables in your data. It would be a big task if you have a very large number of variables, but it will at least show you if there are any variables that appear to have no bearing on what is happening (i.e. their correlation is very low and of no statistical significance). Armed with this knowledge, you will probably not wish to pursue investigating questions that look at those relationships. Not only does the matrix identify correlations and their statistical significance, they also calculate their statistical significance.

2. ***Partial correlation*** compares your *correlation matrix* with a matrix produced after you control for the effect of another variable. In the chapter, the example of an employee salary survey showed how this might work. The data showed a very high correlation between current and starting salary (not surprising, really), but it also showed moderate to high correlations between current salary, gender, educational attainment and job level. So, if there is a very high correlation between current and starting salary, do these other variables have any moderating role? ***Partial correlation*** can help to answer that question. This involves looking at what happens to the correlation between salaries when the effect of each of these is controlled for in turn. This helps us determine what other influences are at work in these relationships. However, it is quite easy to become lost in the quagmire of data and it is important to be systematic about the whole process and ask a question that you can then examine the data for an answer. For example, in the chapter, the questions being asked were:

    - Is the relationship between salaries based purely on these two variables, or are there other controlling variables?
    - The strongest relationship between salaries and other variables is with job level. Is job level a significant other factor?
    - Similarly, educational attainment is the next strongest correlation. Does this influence the relationship between current and start salaries?
    - Finally, does gender have an effect?
    - Which of these controlling variables is having the greatest effect on the salaries variables?

3. ***Multiple regression*** is a procedure based on *linear regression* and is designed to create a mathematical formula with which to predict a value of a *dependent variable* for any known *independent variable* (or vice versa). The procedure effectively calculates formulae to be applied to each of the independent variables.

4. The ***stepwise*** method of *multiple regression* carries out the process by including variables into the equation in order of influence. At each stage, the software checks for statistical significance and once it exceeds the limit set by the researcher (always at the 0.05, 0.01 or 0.001 level), the procedure is halted because it is no longer valid to continue. It often occurs that variables you identified for inclusion in the procedure are not included because their effect is not statistically significant.

5. ***Bivariate tables*** (*cross-tabulations*) are very useful in reporting data in a way that makes the relationships you have identified easy to understand and see. This is especially important where the reader is less likely to be statistically competent. The use of correlation matrices and regression printouts and other technically complex visual material should be reserved to technical appendices or in reports designed to be read by statistically competent people.