Introducing some advanced techniques: multilevel modelling and structural equation modelling

In this final chapter we will briefly cover two techniques which have their roots in regression analysis, but which involve more advanced forms of analysis: multilevel modelling and structural equation modelling. These techniques can be used to ask questions of your data that would not be possible using the standard regression techniques we have described so far. The information presented here should be treated as a very brief overview and for those who wish to pursue the subject, there is further reading that we recommend. These techniques often cannot be carried out with many of the more common statistical analysis packages; at the end of the chapter, we will have a brief look at some of the software that can do these types of analysis.

8.1 Multilevel modelling (MLM)

It is often said that necessity is the mother of invention, and MLM is an example of an analysis technique that was developed for a specific need. In many types of study we collect data from people who are grouped into clusters. In this case measurements from these people will not be independent, and hence we will violate the assumption of independence, discussed in Chapter 4. There, we said that every case should be equally related (or unrelated) to every other case. MLM was developed, and was originally used in school-effectiveness studies. Pupils in schools are clustered into classes, classes are clustered in schools, schools are clustered in areas, areas in districts and districts in countries. Two pupils who are in the same classroom will be more similar to each other than two pupils who are in different classrooms; similarly two classrooms in the same school will be more similar to one another than two classrooms in different schools, and so on. Data that have this type of structure are referred to as multilevel data, or hierarchical data. MLM is sometimes called hierarchical linear modelling (HLM). Some books also refer to ‘random slope modelling’, or ‘mixed modelling’.

Studies in educational effectiveness examine the effects on achievement of factors such as class size, teaching style, or sex segregation. In such a study, we may wish to control for initial ability. In the simplest form, a school-effectiveness study would have one outcome variable (usually examination performance), and two independent variables: a measure of initial ability, taken when the pupils start at the school, and a measure of the variable in which we are interested, for example teaching styles. Deciding upon the most appropriate level of analysis is a problem when using the regression analysis that we have described in the rest of this book to analyse data that have a multilevel structure.

If we were to analyse the data using a usual regression approach, we would analyse individuals, who would each make one case in our analysis. This is usually called the first level, or Level 1. If we were to carry out this analysis, at Level 1, the assumption of independence required by regression models is violated and this violation of the assumption of independence may lead to inflated relationships and a higher type I error rate, as was discussed in Chapter 4. If higher level units are analysed (e.g. classes (Level 2) or schools (Level 3)), then information regarding the actual score of each pupil is discarded. Discarding this information may therefore reduce the measured relationships between variables, and inflate the type II error rate (Goldstein, 1995). A practical example of a type I error was demonstrated by Aitkin et al. (1981). They reanalysed an important and influential study into teaching styles carried out by Bennett (1976). In the original study, which had used a regression-based approach, a significant difference was found between teaching styles. However, when Aitkin et al. reanalysed the data accounting properly for the multilevel structure of the data, the significant differences disappeared.

It is also important to analyse the data at an appropriate level, because the relationship at Level 1 and the relationship at Level 2 may not be the same, and it is very tempting to generalise from one level to another, where this may be inappropriate.

Figure 8.1 shows a (fictional) graphical example of how a failure to account for the multilevel nature of a dataset can lead to a situation where a negative relationship at Level 1 co-exists with a positive relationship at Level 2. When this occurs, generalising from one level to another level can lead to a faulty interpretation, and so we must be especially careful to know which level we are talking about. Students worked on one of three statistical problems, labelled as $w$, $x$ and $y$: the time they spent working on the problem and the grade they achieved were measured. These data are hierarchical, the Level 1 units being the individual students, and the Level 2 units being the problems they worked on. Each of the letters (ignoring for now what the letter actually is) in Figure 8.1 represents a student (Level 1 unit). Each group of letters $(w, x, y)$ represents a Level 2 unit (a problem). The apparent relationship between the two variables, hours and grade, when we analyse these data at Level 1 is represented by the thick black line, which shows a positive relationship—more hours worked leads to higher grades. The cluster of letters surrounding each line represents the hierarchical nature of the data. But you can see that a negative relationship between the number of hours worked and the grade achieved exists within each problem. Thus if we ignore the hierarchical nature of the data, a negative relationship between hours and grade has been disguised, and may be interpreted as a positive relationship. The relationship at Level 1 and the relationship at Level 2 are different, and if we are to talk
However, you will remember from Chapter 1 that this was a simplification. A more general example would be:

$$y_i = b_1 x_{1i} + c + e_i$$

where $y_i$ is the score of the $i$th person on the dependent variable (so where $i = 1$, $y_i$ is the score of the first person), $c$ is the intercept, $b_1$ is the estimate of the slope of the regression line, $x_{1i}$ is the score of the $i$th person on the independent variable, and $e_i$ is the deviation from the predicted score for that person (i.e. the residual).

In simple MLM, we have $i$ Level 1 units (individuals) nested within $j$ Level 2 units (classes). Then, we introduce a new component into the equation, $u_j$. This represents the departure from the constant for each group. Now the equation becomes:

$$y_{ij} = b_1 x_{1ij} + c + u_j + e_i$$

This equation tells us that the score for the $i$th person in the $j$th group is equal to:

- $b_1 x_{1i}$: the slope coefficient, multiplied by the dependent variable; plus
- $c$: a constant; plus
- $u_j$: a departure from the constant for each group; plus
- $e$: a residual for each person.

We can represent this equation graphically, as shown in Figure 8.2, where we use an example that has four Level 2 units (e.g. four different classes). The slope coefficient is equal to the increase in the dependent variable when the independent variable increases by one unit. This slope coefficient is shown in the figure as $b$, which is equal to 1.8. Because all of the slopes are equal we can use any of them to calculate $b$.

The constant is the average point at which the slopes intercept with the $y$-axis. The four different slopes all intercept the $y$-axis at a different point, but the average of all the lines is shown by the thick black line which crosses the $y$-axis at 3; therefore $c = 3$.

The values for $u_j$ are the amount by which the intercepts for each slope differ from the intercept for the average slope. We have four groups, so we have four values for $u_j$. To find $u_1$ we see how far the intercept for Group 1 is above the average intercept ($c$). We can see in the graph that the intercept for Group 1 is equal to 5, since $5 - 3 = 2$, $u_1 = 2$. Similarly $u_2 = 1$, $u_3 = -1$ and $u_4 = -2$.

We can build up the model in a number of different ways. In this model we have forced all groups to have equal slopes, but we can allow slopes to vary across groups. We can also include more independent variables, and these can act at different levels: teaching style would act at the class level, whereas intelligence would act at the individual level. We can cluster the groups into
higher level units, for example schools may be clustered into educational authorities, or districts; slopes can vary within these higher level units and independent variables can also act within them (see Woodhouse (1996) for details).

At higher level units we can examine the relationship between parameters that were estimated in lower level units. Figure 8.3 shows a two-level model where children's achievement was measured when they started at a particular school (Test 1) and then assessed again at a later date (Test 2). The slopes represent each school, and although the slopes differ they are all positive. More importantly, the schools with lower intercepts seem to have steeper slopes; this means that there is a negative correlation between the slope and the intercept (imagine working that out with standard regression analysis!). This shows that in the schools where the children start off as low achievers, the children seem to improve a great deal, whereas in schools where the children start off as high achievers, the children do not seem to improve as much.

8.1.2 Hierarchies everywhere

One of the problems that people find when they discover that multilevel (or hierarchical) data exist, is that suddenly everything becomes an MLM problem. As Kreft and De Leeuw (1998) put it, 'Once you know that hierarchies exist, you see them everywhere' (p. 1). A few examples are described in this section.

Many studies in social psychology involve the analysis of group behaviour. But each of the individuals in each of these groups will likely be more similar to their fellow group members than they are to people in different groups, and data on groups should therefore be treated as hierarchical data. Studies in clinical psychology and medicine often involve the analysis of patients taken from multiple centres – several hospitals, or clinics, for example. Each of these should be treated as a Level 2 unit. Research in occupational psychology can involve the analysis of relationships between variables measured on people who are clustered within departments, departments which are clustered within organisations, and organisations which may be clustered within types of industries. Each of these should be treated as a hierarchy.

8.1.3 Even more hierarchies

Many other types of data structure which could be analysed using 'traditional' techniques can be treated by MLM: the analysis may be equivalent, but the multilevel approach can be much more flexible. For example, in a longitudinal experiment, which would usually be analysed as a repeated-measures MANOVA of some kind, the data on the individuals can be treated as Level 2 data, and the individual measurements from each individual can be treated as Level 1. In the same way that pupils can be clustered within classes, and classes within schools, measures can be clustered within individuals. This approach provides a great deal more flexibility than the usual approach, particularly when measurement of time intervals is being considered. In a MANOVA-based analysis, if a person drops out of the study for any reason that individual usually must be removed from the study. In a study based on the multilevel approach, an individual dropping out of the study is simply...
treated as would be a class with fewer pupils than another class. In addition the length of time the individual did participate in the study can be treated as an independent variable (at Level 2), and can be used as a predictor, in much the same way that a class size can be used as a predictor. A common problem in longitudinal research is determining whether those participants who drop out of a study are a random selection from all participants, or if a variable that we are interested in is causing them to drop out. More importantly, we want to know if the relationship between our variables differs between the participants who drop out and those who remain in the study. Whether participants who drop out are different is relatively easy to determine using a multilevel approach: the number of times an individual participated in the research programme can be treated as an independent variable. By looking for the main effects of this independent variable, we can determine whether the dropouts differ on the outcome variable. If we look for interactions between this independent variable and other variables, we can determine whether relationships between the variables differ for dropouts and non-dropouts.

8.2 Structural equation modelling

Structural equation modelling (SEM) is a general data analytic technique with a number of advantages over traditional approaches to statistical analysis and it is increasing in popularity in psychological research. SEM encompasses everything that ‘traditional’ regression techniques can accomplish, with lots more besides. The effectiveness of SEM is being recognised increasingly by researchers from many disciplines, and the use of SEM techniques has been increasing. Tremblay and Gardner (1996) documented the increasing use of SEM in psychological research and reported that the percentage of articles involving SEM had doubled between 1987 and 1994, and that there had been a similar increase in the number of journals which publish papers involving SEM.

We will first discuss why SEM is used and then introduce some basic concepts that are important to SEM.

8.2.1 Why use SEM?

SEM produces results equivalent to any that regression techniques can produce. In fact, regression analysis can be thought of as one particular type of structural equation model. (You may also remember from Chapter 3 that many other analyses can be thought of as types of regression, and this means that they can also be thought of as types of SEM.) So why, the alert reader asks, have we struggled through seven chapters of a book on regression when we should have learned about SEM from the start? And why have the authors wasted a reasonable proportion of their lives writing about regression when they could have saved the time by telling their readers to go and buy one of the excellent introductions to SEM? We offer two justifications:

1. SEM is difficult. The learning curve involved in SEM is reasonably steep, and most of what we have written in this book can be applied to SEM. To learn about SEM, it is first necessary to have a thorough understanding of regression analysis. It involves many issues which are controversial (and will continue to be controversial for the foreseeable future).

2. To attempt to learn about SEM without this very thorough understanding of regression analysis is not a task that we would care to inflict on anyone.

So if SEM is so difficult and has these unresolved issues, why are we writing about them at all? We have impressed upon the reader throughout this text that regression analysis is a very flexible approach which can be used in a wide variety of situations, and we now say that everything that can be done with regression analysis can be done with SEM. However, there are an even wider range of analyses that SEM analysis can do and regression analysis cannot.

SEM analysis has three main advantages over other forms of analysis and we will examine them in the following sections.

8.2.1.1 Tests of model fit

A regression analysis provides us with a set of parameter estimates and a set of standard errors for those parameter estimates. Because of the nature of the equations for the regression analysis, it is always possible to find satisfactory solutions to the equations. The model that we get from a regression analysis can never be ‘wrong’, in the sense of not fitting the data. (The model may be irritating, unsatisfactory or surprising, it may mean that we need to rethink our ideas about the structure of the processes underlying our measures, but it cannot be wrong.)

When we analyse data using an SEM approach, we formulate a hypothesis about the underlying model and we test that hypothesis. If the model is appropriate, we can interpret the parameter estimates. However, it is possible — even usual — for us to be wrong about the model. If the model is wrong, or (in SEM terminology) if the model does not fit the data, the parameter estimates will not be meaningful, and cannot be interpreted.

Determining whether a model is adequate in terms of fit is a complex problem, often hotly debated, so here we offer only a simple account of model fit.

8.2.1.2 Flexibility

We said in the previous section that SEM is a highly general analysis technique, and being so general makes it suitable for testing the types of complex