ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/256801601

An introduction to statistical modelling

Chapter · January 2011

CITATIONS
7

READS
24,705

1 author:

Kelvyn Jones
University of Bristol
257 PUBLICATIONS 9,765 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Neighbourhood effects View project

Project Multilevel modelling of property(house) prices View project

All content following this page was uploaded by Kelvyn Jones on 26 May 2014.

The user has requested enhancement of the downloaded file.

27

An Introduction to Statistical Modelling

Kelvin Jones, School of Geographical Sciences, University of Bristol, UK

Summary

- Regression modelling
  - Researching ‘cause and effect’ relations that are neither necessary nor sufficient
- Multilevel modelling
  - Researching a specific problem – and how it relates to different forms of multilevel structures.
- Key theorists and writers
  - Paul D. Allison
  - Harvey Goldstein
  - Kelvyn Jones

Statistical modelling is a huge subject. In the space we have available I will concentrate on why you do modelling and what can be achieved. I consider what sort of questions it can answer, what sort of data looks like a ‘regression’ problem and what steps we can take to ensure we get valid results. I have written this introduction from the advanced perspective of the generalized linear model (McCullagh and Nelder, 1989) and have included a substantial discussion on the developing approach of multilevel modelling because of its major potential in the analysis of social research questions.

Key concepts: regression modelling

In the social sciences we research ‘cause and effect’ relations that are neither necessary (the outcome occurs only if the causal factor has operated) nor sufficient (the action of a factor always produces the outcome). Moreover, inherent variation or ‘noise’ may swamp the ‘signal’ and we need quantitative techniques to uncover the underlying patterns to produce credible evidence of a relation (see Jones, Chapter 23, in this volume). A good exemplar comes from epidemiology. There are lung cancer victims who have never smoked, and people who have smoked for a lifetime without a day’s illness. The link was once doubted but we now have unequivocal evidence. Men who smoke increase their risk of death from lung cancer by more than 22 times (a staggering 2,200 per cent higher). The estimate is that one cigarette reduces your life on average by 11 minutes.

To illustrate the arguments I will use a research problem of assessing the evidence for discrimination in legal firms. In that context, statistical modelling provides the following:

- a quantitative assessment of the size of the effect, for example the difference in salary between blacks and whites is £5,000 per annum;
- a quantitative assessment after taking account of other variables; for example a black worker earns £6,500 less after taking account of years of experience; this *conditioning* on other variables distinguishes modelling from ‘testing for differences’ (see Barnes and Lewin, Chapter 26 in this volume);
- a measure of uncertainty for the size of effect; for example we can be 95 per cent confident that the

black–white difference in salary to be found generally in the population from which our sample is drawn, is likely to lie between £4,400 and £5,500 (see Lewin, Chapter 25 in this volume, for an explanation of confidence intervals).

We can use regression modelling in a number of modes; as description (what is the average salary for different ethnic groups?); as part of causal inferences (does being black result in a lower salary?) and in predictive mode ('what happens if' questions). The last can be very difficult to achieve, because change may be so systemic that the underlying relations themselves are altered, and past empirical regularities captured by the modelling no longer hold in a period of regime change (Lucas, 1976).

Data for modelling

Modelling requires a quantifiable outcome measure to assess the effects of discrimination. Table 27.1 provides several, differentiated by the nature of the measurement. There is a continuous measure of salary; the binary categorical outcome of promoted or not; the three-category outcome (promoted, not promoted, not even considered); a count of the number of times rejected for promotion; and a time-to-event measure, the length of time that it has taken to promotion, where a '+' indicates that the event has not yet taken place. All of these outcomes can be analysed in a generalized linear model, but different techniques are required for different scales of measurement. Suitable models going from left to right across the table are normal-theory, logit, multinomial, Poisson, and Cox regression but they all share fundamental characteristics of the general family (Retherford and Choe, 1993). Also shown in the table are a number of 'explanatory' or predictor variables, again with different scales of measurement. Gender is measured as two categories, ethnicity as four, education as a set of ordered categories, and years of employment on a continuous scale. All of these scales can be analysed in the general framework.

Relations

Figure 27.1 displays a range of relations between a response, salary, on the vertical axis and predictor variables on the horizontal. In (a) there is a sizeable difference between the male and female average income. In (b) to (d) we see a number of straight-line relations between salary and years of service. The first is a positive one; the longer you have worked for the firm, the more money you get. The second is the flat one of no relation; there is no effect of length of employment on pay (think fast-food outlets!). The third shows a negative relation, the longer you have been there, the less you get paid (this can happen in physically demanding jobs).

A non-linear relation between salary and length of employment is shown in (e); an initial steep rise tails off indicating that the full salary is reached rapidly. In (f) salary increments get steeper and steeper with experience, and in (g) there is a curvilinear relation such that salary increases for the first six years then tails away. An interaction between gender and length-of-employment is shown in (h). At appointment there is no gender gap but this opens up the longer you are employed. The distinctive feature of (i) is that in addition to the solid lines displaying averages for the four categories of ethnicity, there are dashed lines representing the confidence interval (see Lewin, in this volume). We can be 95 per cent confident that the true population value will fall within this interval given our sample data. Here, the average white salary is estimated with the greatest reliability and has the narrowest band. The Asian band is the widest; we are unsure what the average for this group is. While the black salary is unequivocally lower than the white as the confidence intervals do not overlap, the evidence is not sufficient to decide on white–Asian differences, nor on Asian-black differences. The unknown group looks indistinguishable from the white group, with a slightly wider confidence interval. The final graph (j) is a *three-way interaction* between gender, ethnicity, and length of employment. At the outset, there are substantial differences between the groups, and as time proceeds, black Women would appear to be doubly discriminated against.

Conditioning

We may be interested in the effect of just one variable (gender) on another (salary) but we need to take account of other variables as they may compromise the results. We can recognize three distinct cases:

- Inflation of a relation when not taking into account extraneous variables: a substantial gender effect could be reduced after taking account of ethnicity; this is because the female labour force is predominantly non-white and it is this group that is characterized by poor pay.

Table 27.1 A dataframe for regression modelling for the discrimination study

	Responses					*Predictors*			
Respondent number	*Salary (£k)*	*Promotion (2 category)*	*Promotion (3 category)*	*Number of rejections*	*Time to promotion (yrs)*	*Gender*	*Ethnicity*	*Years of education*	*Years of service*
1	32.4	No	No	1	6.2+	Female	White	<11	9.1
2	40.1	Yes	Yes	0	3.2	Male	White	11-13	6.2
3	65.2	Yes	Yes	0	2.9	Male	Asian	14-16	4.9
4	32.1	No	No	2	8.2+	Female	Black	>16	8.2
5	21.6	No	Not	4	6.7+	Female	Unknown	11-13	6.7
6	25.4	No	Not	3	4.2+	Male	Black	<11	4.2
7	32.7	No	No	1	5.1+	Female	White	14-16	5.1
8	51.7	Yes	Yes	0	3.9	Male	White	<11	4.8
9	44.0	Yes	Yes	0	4.2	Female	Asian	14-16	7.2
10	32.6	No	No	1	3.9+	Female	Black	14-16	3.9
11	41.7	Yes	Yes	0	4.9	Male	White	11-13	9.7
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
500	39.7	No	No	2	5.2	Male	Unknown	14-16	8.1

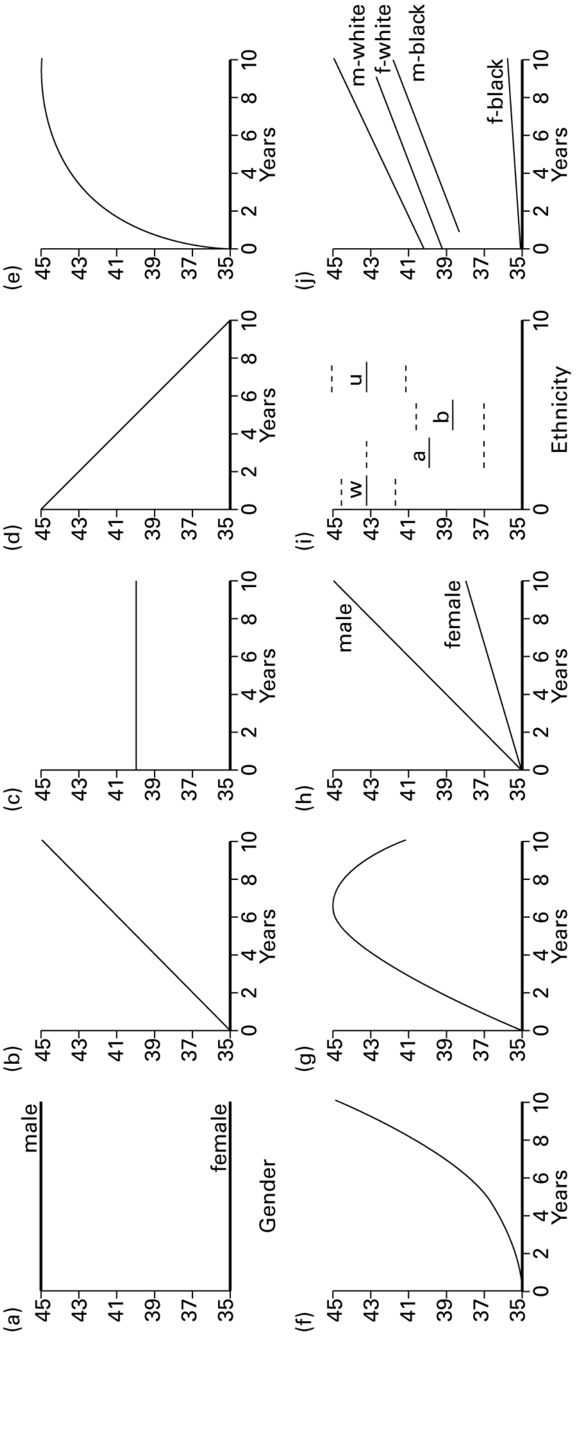

Figure 27.1 Relations between variables

- Suppression of a relation: an apparent small gender gap could increase when account is taken of years of employment; women having longer service and poorer pay.
- No confounding: the original relation remains substantially un-altered when account is taken of other variables.

While modelling can usually assess the partial relationship between two variables taking account of others, this cannot be achieved when predictor variables are so highly correlated that we have no effective way of telling them apart. In the pathological case of exact collinearity (complete dependence between a pair or more variables) a separate effect cannot be estimated. For example, if all Asians in the survey are women, we cannot determine the gender gap for Asians. More generally, collinearity is a matter of degree and as the correlation between predictor variables increases, so do the confidence intervals as there is insufficient distinctive information for reliable estimation.

Form of the model

All statistical models have a common form:

Response = Systematic part + Random part

The systematic part is the average relation between the response and the predictors, while the random part is the variation in the response after taking account of all the included predictors. Figure 27.2 displays the values representing the data for 16 respondents, and a straight line we have threaded through the points to represent the systematic relation between salary and length of employment. The line represents fitted values; if you have 10 years service you are predicted to have a salary of about £45,000.

All equations of the straight line involving two variables have the same form:

Fitted value = Intercept + (Slope*Predictor)

which here is:

Predicted salary = Intercept +
(Slope*Years of service)

which (say) we estimate to be:

Predicted salary = 30.3 + (1.7*Length of service)

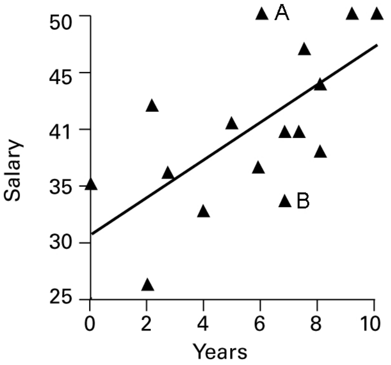

Figure 27.2 Simple linear regression

The intercept gives the predicted value of the response when the predictor takes on the value of zero, so we are predicting that the average salary on appointment (when years of service is 0) is £30,000. The slope gives the marginal change in the response variable for a unit change in the predictor variable. For every extra year with the firm, salary increases by £1,700. Importantly, the increase in salary consequent from staying 0 to 1 years in service is the same as from 5 to 6 years of service. This is a direct consequence of assuming that the underlying functional form of the model is linear and fitting a linear equation.

The term random means ‘allowed to vary’ and, in relation to Figure 27.2, the random part is the variation in salary that is not accounted for by the underlying average relationship with years. Some people are paid more and some less given the time they have been with the firm. We see that person A has an income above the line while B is below. The difference between the actual and predicted salary is known as the residual. In fitting the line we have minimized these residuals so that the line goes through the middle of the data points. Here we have used a technique called Ordinary Least Squares in which the sum of the squared residuals is minimized. Responses with other scales of measurement require other techniques, but all of them are based on the same underlying principle of minimizing the ‘poorness of fit’ between the actual data points and the fitted line.

In some cases there will be a close fit between the actual and fitted values but in other cases there may be a lot of ‘noise’, so that for any given length of employment there is a wide range of salaries. It is helpful to characterize this residual variability. To do so requires us to make some assumptions. We need to conceive of the residuals as coming from a particular distribution. Given that salary is a continuous variable, we can assume that the residuals come from a Normal distribution (other scales would suggest other distributions). If we further assume that there is the same variability for short and long length of service (that is homoscedasticity) we can summarize the variability in a single statistic, the standard deviation of the residuals. For Figure 27.2 this value is 6, and we can anticipate (given the known properties of a Normal distribution) that the income for 95 per cent of employees on appointment will lay roughly 2 standard deviations around the mean value of £30,000. Most people will have an initial income between £18,000 and £42,000. This rather wide spread of values is due to inherent uncertainty in the system and the small number (16) of observations we have used. Another key summary statistic is the R-squared value which gives the correspondence between actual and fitted values, on a scale between zero (no correspondence) and 100 (complete correspondence). Chapter 23 warns about the over-reliance on such goodness of fit statistics, and you are encouraged to undertake a residual analysis to diagnose problems with the fitted model; a good account can be found at: http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

The model we have so far discussed is a 'simple' one with only one predictor. In a multiple regression model there is more than one predictor (there can be any combination of continuous and categorical variables) with the following differences:

- *Intercept*: this is now the average value for the response when all the predictors take the value zero;
- *Slope*: there is one for each predictor and this summarizes the conditional or partial relationship as the change in the response for a unit change in a particular predictor, holding all the other predictors constant;
- *Residual*: the difference between the actual and fitted values based on all the predictors;
- *R-squared*: the percentage of the total variation of the response variable that is accounted for by all predictors taken simultaneously.

Figure 27.1 (h) to (j) are all examples of multiple regression models, the key to their specification being the coding of predictor variables. A comprehensive discussion of how to do this can be found at: http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter5/

Implications for research design: regression modelling

We can recognize two broad classes of design that will produce suitable data for regression modelling; experiments where we intervene and observational studies. Experiments are artificial settings in which we change the predictor variable and see what happens to the response, while keeping other variables 'controlled'. We can also randomly allocate each case to an 'intervention' or not, thereby guaranteeing that any detectable change in the response is due to the intervention. Because of this closure and control of unknown factors, experiments are a very strong procedure for causal inference. It is often thought that experiments are more or less impossible in most social sciences due to ethics and relations being disposition-response, not stimulus-response. You cannot easily change a person's gender and keep everything else the same! But with some ingenuity we could get something like the data we need. If we are interested in how ethnicity affects whether a person is promoted or not, we could write scripts for an interview, varying some elements such as length of employment, but keeping all the rest the same. Actors of different ethnicity could record these, and the videos would be played to managers to see what decision they would come to. Modelling would then identify the size of the effect of ethnicity in relation to years of employment. This is a very strong design for causal inference but the external validity may be weak due to the artificiality of the process, so that everyone gets promoted!

Observational designs are less strong for causal inference, but if attention is paid to scientific sampling so that each member of the population has a known chance of inclusion, they can be highly representative. We can recognize four broad groups of design, each with their own strengths and weaknesses: administrative data, cross-sectional surveys, case-comparison study and panel design (see Jones and Moon, 1987). All of these designs can yield data that can be modelled by regression analysis, the choice of design being determined by the type of question being asked and the resources available. There is one golden rule that must be followed however: 'the specifics of the design must be taken account of in the modelling'. For example, a panel survey (where people are tracked periodically) will generate data for respondents that will be patterned across time (salary now will be similar to what it was last year and the year before), and this 'non-independence' must be explicitly modelled.

Our aim in designing how we are going to collect the data and how we are going to analyse them is to get valid results. We can recognize two broad areas of validity that particularly apply to the analysis and the design of a model-based study, and we will discuss these issues using regression in causal mode.[1]

Conclusion validity

This is concerned with analysis, and asks if the conclusions we have reached about relationships in our data are credible. We can be wrong in two ways: missing a real relation; finding a relation where there is none.

The key threats to this sort of validity (and what to do about them) are.

- The assumptions of the systematic part (e.g. in terms of linearity) and the random part (in terms of the nature of the distribution and such properties as homoscedasticity, that is equal variance and independence)[2] must be met. This amounts to the systematic part of the model fully capturing the generalities of the world; equivalently the random part is just 'trendless' fluctuations. We can use 'diagnostics' to assess assumptions and robust procedures with less demanding assumptions. A useful guide to both these approaches is to be found in Cook and Weisberg (1999).
- Data dredging: This is analysing the data repeatedly under slightly differing conditions or assumptions, dropping these cases, transforming this variable, trying out a very large number of different predictor variables, or including every possible interaction to maximize the R-squared. If we do this, we are more or less bound to find something. But the status of what we have found is problematic; we cannot tell whether what we have found is idiosyncratic noise or generalizable signal. The best advice is to focus on a single topic. We should ask not the vague 'what determines salary', but 'is there discrimination by ethnicity in annual salary when account is taken of gender and length of employment?' If you do undertake such model searching, keep it limited, be honest in your write-up, adjust your level of significance to take account of multiple hypothesis testing, and use a hold-out sample as an independent test of the model.
- Lack of statistical power so that the sample is too small a sample to detect a real relationship: The required number of observations is determined by three factors: noisy systems need more observations, so do predictor variables lacking variability, and collinear predictors. As a very rough rule of thumb, you would not usually have more than 10 predictor variables in a single model, and you might plan on collecting at least 25 observations for each. Software is available (e.g. http://www.insp.mx/dinf/stat . . . list.html) which indicates required sample size for a given power. A common rule of thumb is a power of 0.8; at least 80 percent chance of finding a relationship when there is one.
- Measurement error: We can have imprecise measurements and we can have systematically biased measurements. In general, biased measurements will produce biased estimates of effects, unless all variables are off target by the same amount. Non-systematic errors in the dependent variable will require additional observations for the same power; while such errors in the key predictor variable usually biases estimates by attenuating them to zero. We can do a sensitivity analysis to appreciate the effects of measurement error, while during collection we can use a pilot survey to assess reliability and bias. Developing a consistent protocol, training the interviewers and careful wording of questions can all help.

Internal validity

This second type of validity addresses the question of whether the relationship we have found is a causal one. The key threats to validity (and what to do about them) are:

- *Omitted variables bias* refers to alternative explanation of results: to be problematic, such variables must be related to *both* the response and the included predictor variable. Specification error tests are available (Hendry, 2000), but while these may indicate a problem, they cannot suggest what variable is missing. This is the Achilles heel of regression modelling with observational data; with

j:tmsrply 2-8-2010 p:248 c:0

an experiment employing randomization this should not be a problem. The best possible advice is to think hard about a research problem and include all relevant variables. At the same time you do not want to include irrelevant variables, as this will reduce the power of the design to detect real effects. It can help a great deal to classify possible predictors into direct causes, indirect causes, moderating variables and mediating variables (Miles and Shevlin, 2001).

- *Endogeneity* is a fancy term for having a predictor variable that is directly influenced by the response, such that income is determined by health and health by income. In an experiment, this problem is ruled out by design as you can manipulate the predictors and see the subsequent effects. With observational designs there are specialist techniques such as instrumental variables for improved estimation of causal effects (Angrist and Pischke, 2009). Panel designs can also be vital here. In some situations it is possible to rule out this problem a priori; it is unlikely for example that gender or ethnicity are determined by salary!
- *Selection bias* is when we have selected our respondents so as to in some way systematically distort the relation between the predictor and the outcome. The problem is such that any selection rule correlated with the response variable will attenuate estimates of an effect towards zero. For example if we had only been able to collect data on those above an income threshold, we would have attenuated the relation between salary and years of experience. People who do not return your questionnaire may be different, in some important way, to the people who did. Strict adherence to sampling protocols, well trained interviewers, and intensive follow-up to a pilot can help minimize this problem. There are also analytical techniques that can adjust the estimates to take account of this bias.

It is worth stressing in concluding this section what regression modelling is trying to do. It aims at generality and generalizable results. We are not primarily interested in why this specific person did or did not get a salary rise, but what is happening to females as a group. We can only collect sample data but we wish to infer quite generally what is going on across the country. Once identified, this generality throws into stronger relief any unusual cases. We are continually searching for evidence that supports/challenges alternative explanations and we are always looking for the empirical implications of our theory to subject it to rigorous evaluation.

Key concepts: multilevel modelling

Multilevel modelling is a recently developed procedure that is now seeing widespread use. It is given a separate section here because of its potential for handling a wide variety of research designs. Although it grows out of regression, the approach represents a considerable increase in sophistication. We begin with a specific problem, and then show how this relates to different forms of multilevel structures and associated research designs.

A multilevel problem

My university (in the UK) like others has been keen to widen its participation. It may be that if an able student goes to a poorly performing school, their A level score at entry is an under estimate of their potential.[3] Alternatively, if they go to a fee-paying, highly-resourced school, their score has been temporarily boosted, and this does not carry over to their degree performance. If we can identify such situations we may justifiably recruit students with a lower point score on the basis of greater potential. But what is the evidence for such a policy? We can set this up as a regression-type problem in which the response of the degree result of the student is related to three predictors: A-level score, the school average performance and an indicator of school type.

But there is a difficulty because we are dealing with a problem with a multilevel structure. Student and school are not at the same level in that (many) students are nested in (fewer) schools. Moreover, students belonging to the same school are more likely to be alike than students from different schools. If this 'non-independence' is not taken into account, we have fewer observations than we think we have and we run the risk of finding significant relationships where none exist. Technically the effective degrees of freedom (see Barnes and Lewin, in this volume) are lower than we think they are. But this is more than a technical problem, for there are several sources of variation that need to be taken into account for a proper analysis. Thus, there is the between-student variation, between-school variation and, extending the analysis, between-university, and between-discipline variation. In relation to the last there may be disciplines where the A level score is a very poor guide to degree performance, and should not be used as the main entry requirement. Thus the effect of an A level score on performance is not fixed but varies from context to context, where context is provided by the different levels in the structure. In comparison to standard regression models, multilevel models have a more complex random part.

Research designs and multilevel structures

It turns out that a very large array of research questions can be seen as combinations of just three types of multilevel structure that can now be routinely handled by computer-intensive procedures. The simplest structure is the hierarchy in which a lower-level unit nests in only one higher-level unit (Figure 27.3). The classic example (a) is the two-level model in which pupils are nested in schools. This can readily be extended so that pupils at level 1 can be nested within classes (level 2) within schools (level 3) within local education authorities (level 4). This strict hierarchy includes a number of research designs that you might not initially conceive as multilevel problems. A panel design is shown in 27.3(b) where repeated measures (at level 1) are elicited for voting behaviour for individuals (level 2) who are nested in constituencies (level 3). In 27.3(c), there is a multivariate design in which three responses measuring health-related behaviour (at level 1) are nested within individuals (level 2) within places (level 3); the responses are seen as repeated measurements of individuals, and individuals are repeated measures of places. Other examples with such a hierarchical structure include: an experimental design in which the intervention is not made for individuals (at level 1) but for communities (level 2); and an observational design in which there is a two-stage sampling process, first areas (which then become level 2), and then respondents within them (at level 1).

The other types of multilevel structure are two different non-hierarchical structures. The classic example (Figure 27.4(a)) of a cross-classification is students (level 1) being nested within neighbourhoods and also schools (both at level 2). Not all the students in a neighbourhood go to a particular school and a school draws its pupils from more

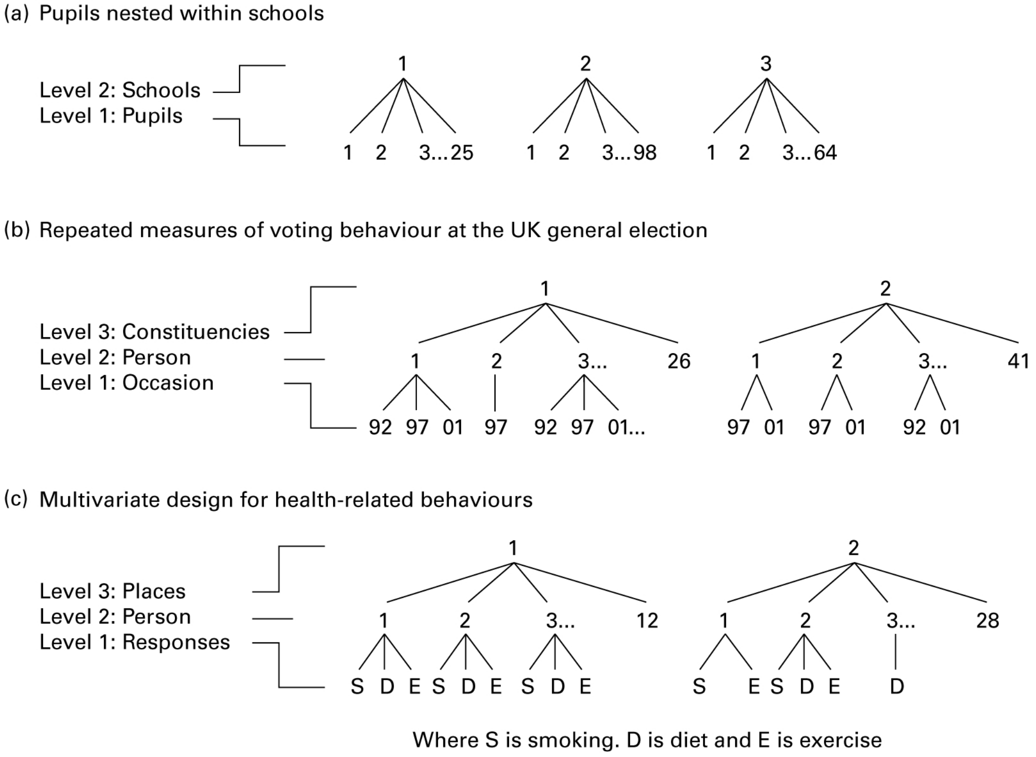

Figure 27.3 Hierarchical structures as unit diagrams

than one neighbourhood. Thus schools and neighbourhoods are not nested but crossed. The final structure is the multiple membership in which a lower level unit 'belongs' to more than one higher level unit (Figure 27.4(b)). Thus a student (at level 1) may be nested within teachers (level 2) but each student may be taught by more than one teacher. We might include in the analysis a 'weight' to reflect the proportion of time each pupil spends with a teacher, so that student 2 spends 50 percent of their time with teacher 1 and 50 percent with teacher 2. Again a large number of problems can be cast within this framework, for example a dynamic household study in which individuals 'belong' to more than one household over time. A less obvious example is a spatial model in which individuals are affected by the neighbourhood in which they live and also by surrounding neighbourhoods – the weight in the multiple-membership structure being some function of distance from the home neighbourhood to the surrounding neighbourhoods. These models can be extended to look at pupil achievement in situations where there is 'competition' between the higher level units, such as schools with overlapping catchments, perhaps differentiated by school types.

An alternative way of conceiving and visualizing structures is as classifications. A 'classification diagram' is particularly helpful for complex problems. Figure 27.5 shows some examples of hierarchical and non-hierarchical structures using this type of diagram: (a) is a 3-level hierarchical problem; (b) is a cross-classified design; (c) is a multiple membership structure; and (d) shows a spatial structure. Boxes represent each classification with arrows representing nesting, single arrows for single membership and double arrows for multiple membership. Returning to the student performance example, we can see it as a combination of these three types of structure (Figure 27.5(e)). Students are nested within schools, and students are nested within disciplines within universities. Schools and universities are crossed because not all the students from a school go to one university. While the student/school relation might be conceived as a strict hierarchy (the last school attended) the university/discipline structure can be seen as a multiple membership one, in that students move between subjects and universities after starting courses.

Importance of structures

Is all this realism and complexity necessary? It is important to realize that a simple model tells you little about a more complex model, but a more complex model provides information about the simpler models embedded within it. Such complexity is not being sought for its own sake, but if the real world operates like this, then a simpler under-specified model can lead to inferential error.

There are in fact two key aspects of statistical complexity. We have so far concentrated on *dependencies arising from structures*. Once groupings are established, even if their establishment is random, they will

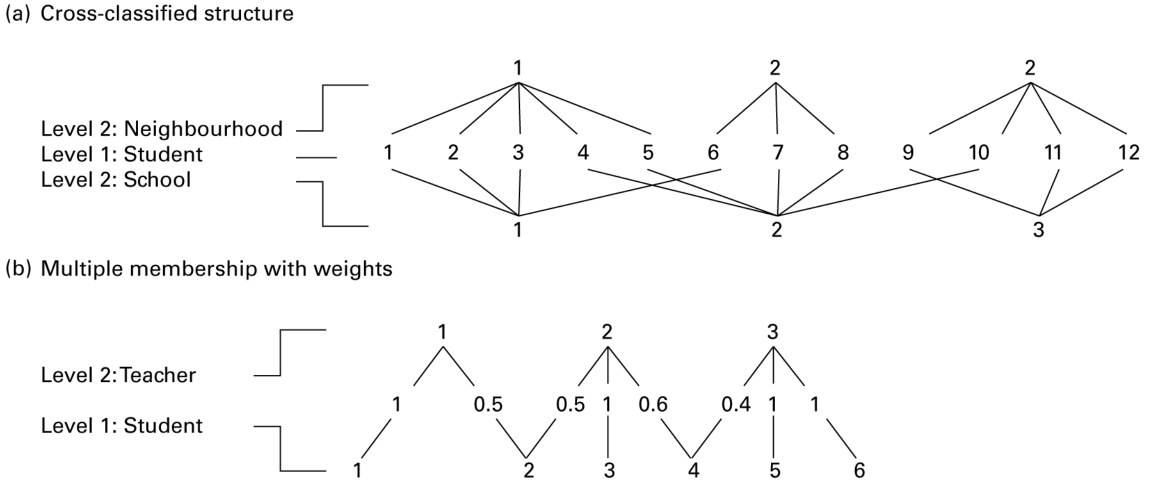

Figure 27.4 Non-hierarchical structures as unit diagrams

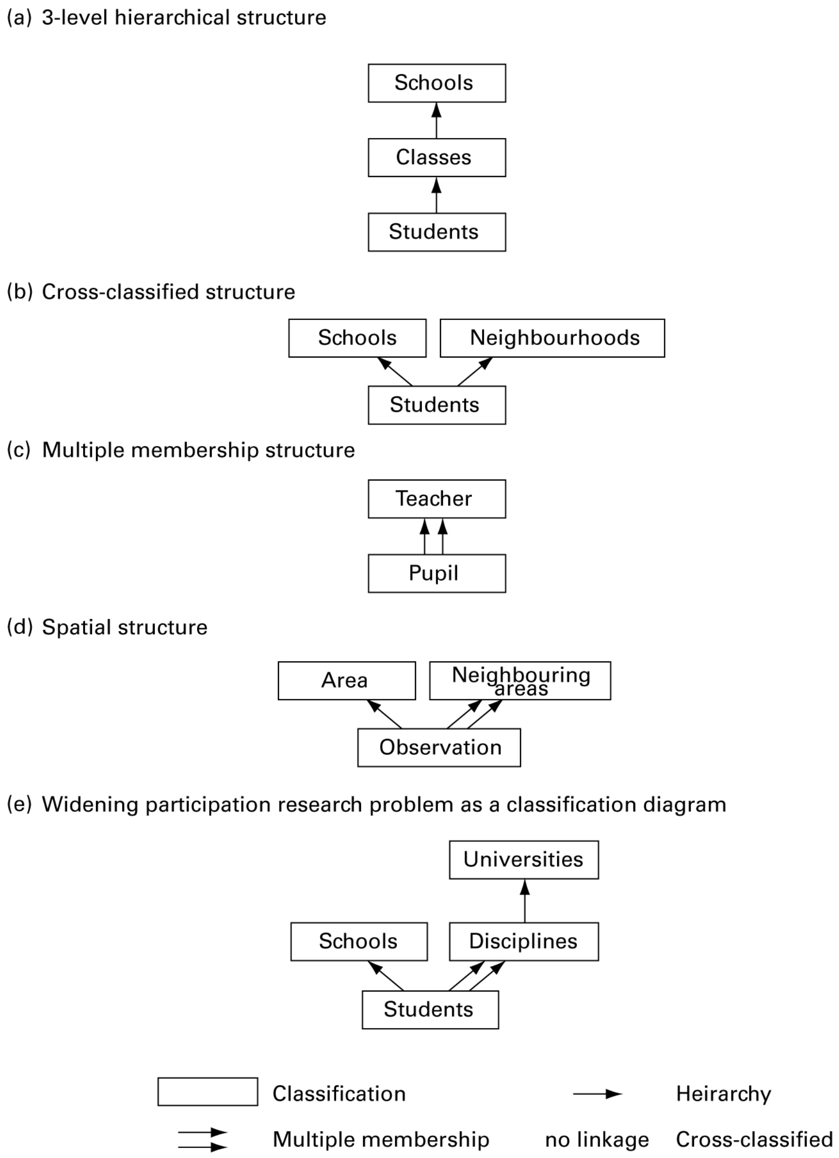

Figure 27.5 Structures as classification diagrams

tend to become differentiated as people are influenced by the group membership. To ignore this relationship risks overlooking the importance of group effects, and may also render invalid many of the traditional statistical analysis techniques used for studying relationships. An example is Aitkin et al.'s (1981) re-analysis of the 'teaching styles' study. The original analysis had suggested that children's academic achievement was higher if a 'formal' teacher using all-class activities taught them. When the structure of children into classes was taken into account, the significant differences disappeared and 'formally' taught children could not be shown to differ from the others. Some data, such as repeated measures and individuals within households, can be expected to be highly dependent, and it is essential this dependency is taken fully into account.

The second aspect is *complexity arising from the* *measurement process* such as having 'missing' data or having multiple measuring instruments. In an observational study we can expect that there will be a different number of pupils measured at each school as shown in the unit diagram of Figure 27.3(a). In a panel study, each person may not respond every year, while in the multivariate design not all people respond to all questions that form the outcome variables. A defining case of the latter is the matrix sample design where all students are asked a core set of questions on say mathematics but different random subsets of pupils are asked detailed questions on either trigonometry, algebra, or set theory. Treating this design as a hierarchical structure allows the analysis of the full set of data in an overall model.

Prior to the development of the multilevel approach, the analyst was faced with mis-applying single-level models, either aggregating to the single level of the school and risking the ecological fallacy of transferring aggregate results to individuals, or working only at the pupil level and committing the atomistic fallacy of ignoring context. The standard model is mainly concerned with averages, and the general effect, where reality is often heterogeneous and complex. Thus females may not only perform better than males in terms of degree results, but they may also be more consistent (more homogenous) in their performance. It is this analysis of structures, contextual effects, and heterogeneity that is tackled by multilevel models.

Figure 27.6 portrays graphically some elements of the widening participation problem in terms of this

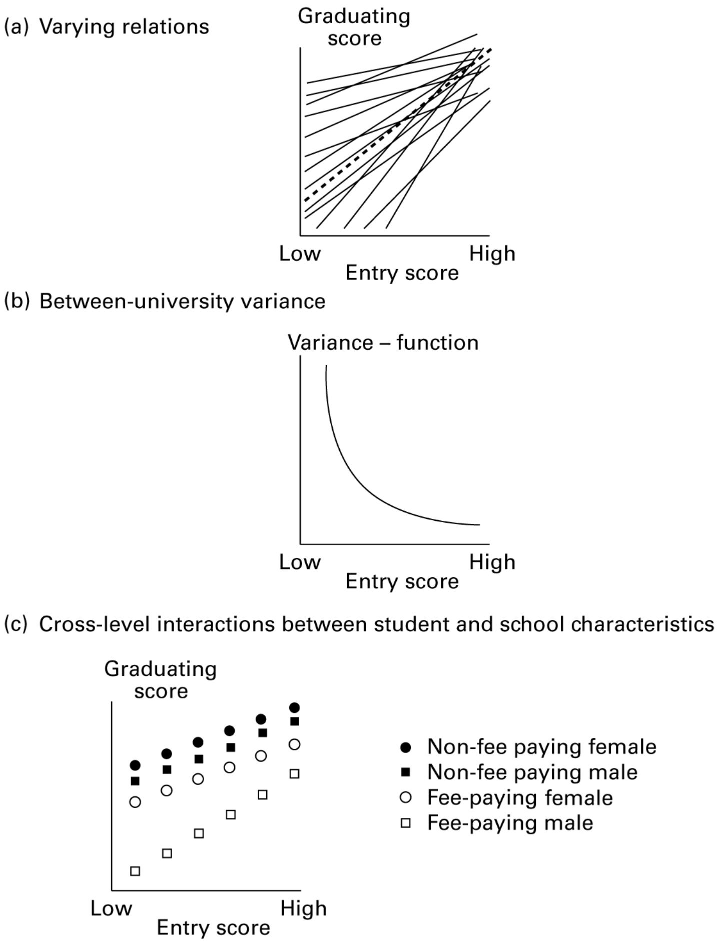

Figure 27.6 Achievement varying over context

heterogeneity and contextuality. In (a), the vertical axis is the final year points score of graduating university students, while the horizontal axis is the points score on entry. The lines shown are the different sampled universities. There is a noticeable 'fanning-in' so that highly qualified students on entry achieve the same excellence irrespective of where they study. But for those with a low-score at initial entry, it makes a great deal of difference where they study. This varying relation between pre-score and post-score is shown in an alternative form as a variance function in Figure 27.6(b) with the same horizontal axis, but the vertical axis is now between-university variance. As the pre-entry score increases, the between-university variation decreases. Finally Figure 27.6(c) shows what is known as a cross-level interaction; the axes are the same as (a), but the four lines represent fee-paying and non-fee paying schools for male and female students. The cross-level interaction is between a school-level variable (fee-paying or not) and two individual-level variables (gender and pre-entry score). The noticeable result in terms of our research question is that students who have attended fee-paying schools do less well across the entry range, but this is most marked for males, especially those with relatively low entry scores. If these results were confirmed, a university may be justified in taking students with lower entry grades from non-fee paying schools, particularly for males if it wished to pursue a policy of equal opportunity.

Implications for research design: multilevel modelling

A major difference between the design of multilevel and single-level studies is the requirements for sufficient power to detect effects. It is not the overall number of observations but the number at each level that is important. Advice depends on the amount of underlying variation in the 'system' being modelled, and what are the main aims of the analysis. Thus, in planning a school-effects study an absolute minimum would be 25 pupils in each of 25 schools, preferably 100 schools. Any less of the higher units will give poor estimates of the between school differences, particularly if school effects are being examined on several dimensions, for example in relation to high and low ability pupils. At level 1 the number of pupils within a school is an important determinant of what can be reliably inferred about a particular school. Little can be said about a particular school if only few students have been sampled. In contrast, if the higher-level unit is a household, there would be very few containing 25 individuals! But this is not a problem for we are unlikely to want to infer to a named household. Instead, we want to know about between-household variability in general, and that is determined by the number of households, and not by the number of people in a household. Finally if we only sample one person in each household we would be totally unable to separate household effects from individual effects. Hox (2002) provides an accessible discussion of statistical power in multilevel models.

Conclusion

Social reality is complex and structured. Recent developments in multilevel models provide a formal framework of analysis whose complexity of structure matches that of the system being studied. Modern software allows the estimation of very complex problems with multiple levels of nesting, and many units such as hundreds of thousands of students. Some examples of this approach are given in Jones (Chapter 29, in this volume); the usefulness of multilevel models in reality in addressing the widening participation issue can be seen from a study entitled 'Schooling effects on higher Education Achievement' available at http://www.hefce.ac.uk/pubs/hefce/, while Raudenbush and Bryk (2001) provide extensive discussion of modelling school effects.

Notes

1. Space precludes a discussion of construct validity (the extent to which variables faithfully measure concepts) and external validity (the extent to which we are able to generalize from our study).
2. To estimate correctly the confidence intervals for an effect requires that the residuals are independent; that is, knowing the value of one should tell you nothing about the value of another.
3. A' levels are public examinations taken in the final year of secondary education; usually at 18 years; good scores are normally required to secure a university place.

Annotated bibliography

Regression modelling

Allison, P. D. (1999) *Multiple Regression: A Primer*. Thousand Oaks, CA: Pine Forge Press.
An excellent primer which introduces the underlying concepts with a minimum of algebra while covering a wide range of models and their motivation.

Angrist, J.D. and Pischke, J-F. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.
This spells out under what conditions quantitative results can have a causal interpretation.

Cook, R.D. and Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*. New York: Wiley Interscience.
This stresses regression as conditional modelling and diagnostics, which are implemented in their freely available ARC software: http://www.stat.umn.edu/arc/

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.
Provides a comprehensive applied introduction that will take you a long way in appreciating the current practice and excitement of statistical modelling; highly recommended.

Retherford, R.D. and Choe, M.K. (1993) *Statistical Models for Causal Analysis*. New York: Wiley.
A very approachable account of generalized linear models in which the response can be categorical or timed to an event. The title is misleading, however, as there is little discussion of causality.

Tacq, J. (1997) *Multivariate Analysis Techniques in Social Science Research: From Problem to Analysis*. London: Sage.
Discusses quantitative analysis in the context of specific research problems.

Trochim, W. (2000) *The Research Methods Knowledge Base*. Cincinnati: Atomic Dog Publishing. Available at: http://www.socialresearchmethods.net/kb/
Available in printed form or as a website, this 'knowledge-base' provides lots of useful advice on minimizing threats to validity in observational and experimental studies.

Venables, W. and Ripley, B. (2002) *Modern Applied Statistics*. New York: SpringerVerlag.
One of the most comprehensive and up-to-date accounts of all sorts of developments in regression-like modelling with substantial online resources: http://www.stats.ox.ac.uk/pub/MASS4/. It can be used in the 'free' Open-source R software environment: http://www.r-project.org/

Wright, D.B. and London, K. (2009) *Modern Regression Techniques Using R*. London: Sage.
A very broad-ranging account achieved with brevity, and provides computer code.

Multilevel modelling

Hox, J. (2002) *Multilevel Analysis: Techniques and Applications.* New Jersey: Lawrence Erlbaum Associates.
Provides a well-written and approachable introduction to multilevel modelling.

Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Arnold.
This is the definitive (but rather demanding) text. The author's team maintains a comprehensive website at: http://www.cmm.bristol.ac.uk/ including the Lemma online course which is a thorough and well-paced introduction.

Raudenbush, S.W. and Bryk, A.S. (2001) *Hierarchical Linear Models*, 2nd edn. Newbury Park: Sage.
Written by two American pioneers of multilevel modelling, this provides a detailed treatment that is linked to their HLM software.

Singer, Judith D. and Willets, J.B. (2003) *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence*. Oxford: Oxford University Press.
A very gradual account that shows in detail how the multilevel model can be used in the analysis of repeated measures; their worked examples in a number of different software packages (and much else) are provided at: http://www.ats.ucla.edu/stat/examples/

Further references

Aitkin, M., Anderson, D. and Hinde, J. (1981) 'Statistical modelling of data on teaching styles (with discussion)', *Journal of the Royal Statistical Society, Series A*, 144: 148–61.

Hendry, D.F. (2000) *Econometrics: Alchemy or Science*. Milton Keynes: Open University Press.

Jones, K. and Moon, G. (1987) *Health, Disease and Society*. London: Routledge.

Lucas, R.E. (1976) 'Econometric policy evaluation: A critique', in K. Brunner and A. H. Meltzer (eds) *The Phillips Curve and Labor Markets*. Carnegie-Rochester Conference Series on Public Policy, 1: 19–46.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman & Hall.

Miles, J. and Shevlin, M. (2001) *Applied Regression and Correlation Analysis in Psychology: A Student's Guide*. London: Sage.

View publication stats