Mathematics Learning Support Centre

Statistics: Multilevel modelling

Richard Buxton. 2008.

1 Introduction

Multilevel modelling is an approach that can be used to handle *clustered* or *grouped* data.

Suppose we are trying to discover some of the factors that affect a child's academic attainment in English at age 16. The sample of pupils involved in our study will be taught in classes, within schools. We'll probably be interested in the effect of a mix of pupil level factors - e.g.the socioeconomic status of the child's parents, class level factors - e.g. the use of streamed vs unstreamed teaching, and school level factors - e.g. single-sex vs mixed. Multi-level modelling provides a useful framework for thinking about problems with this type of hierarchical structure.

Even if we are mainly interested in pupil level factors, we'll still need to take account of the clustering in our sample. For example, the attainment levels of two children in the same class will tend to be more similar than the levels of two children in different classes. If we use statistical techniques that ignore the clustering - e.g. multiple regression - the standard errors and confidence intervals that we obtain will be unrealistic and we may well conclude that there are real effects, when we are simply looking at random variation.

Multilevel modelling can also be used to analyse repeated measures data. For example, if we are measuring the blood pressure of a group of patients at weekly intervals, we can think of the successive measurements as grouped within the individual subjects. One advantage of the multilevel modelling approach is that it can deal with data in which the times of the measurements vary from subject to subject.

The aim of this handout is to introduce the idea of multilevel modelling. If you feel that this approach is likely to be helpful to you, you may well find it useful to look at some of the material listed in Section 7.

2 Example

We illustrate the idea of multilevel modelling with a set of repeated measures data giving growth patterns for a sample of 26 boys in Oxford, England[1] - see Figure 1. The height of each boy is measured on nine different occasions. The plots for the individual boys run from the bottom left to the top right and are arranged in order of the maximum height

[1]Source of data: Pinheiro and Bates (2000)

attained over the period of measurement.

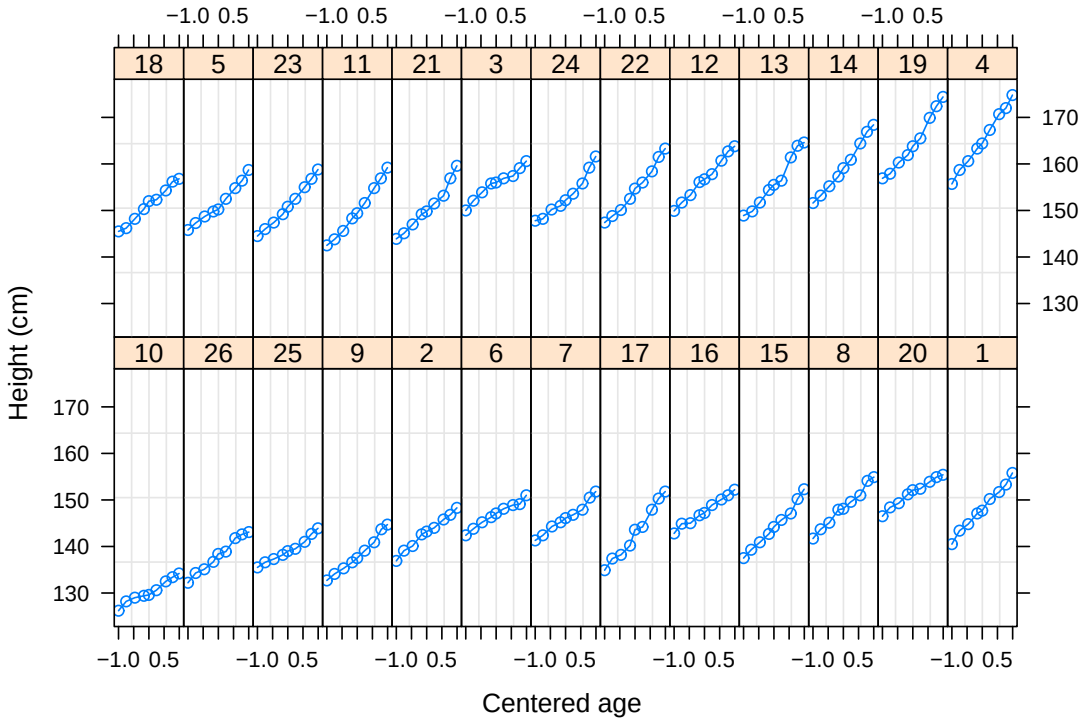

Figure 1: Growth patterns of boys in Oxford

For each boy, the growth pattern appears to be roughly linear, so we might try modelling it by a simple linear regression model of the form...

$$H = \beta_0 + \beta_1 A + \epsilon \tag{1}$$

...where H and A represent height and age and ϵ represents the variation in height that cannot be explained by the linear relationship with age.

To extend our model beyond a single boy, we need to allow for the variation in growth patterns among different subjects. For example, a quick glance at Figure 1 shows that some of the subjects are consistently taller than others - compare subject 4 in the top right of the plot with subject 10 in the bottom left. If we try to use Model 1 for the complete set of data, the fit will be very poor - see Figure 2.

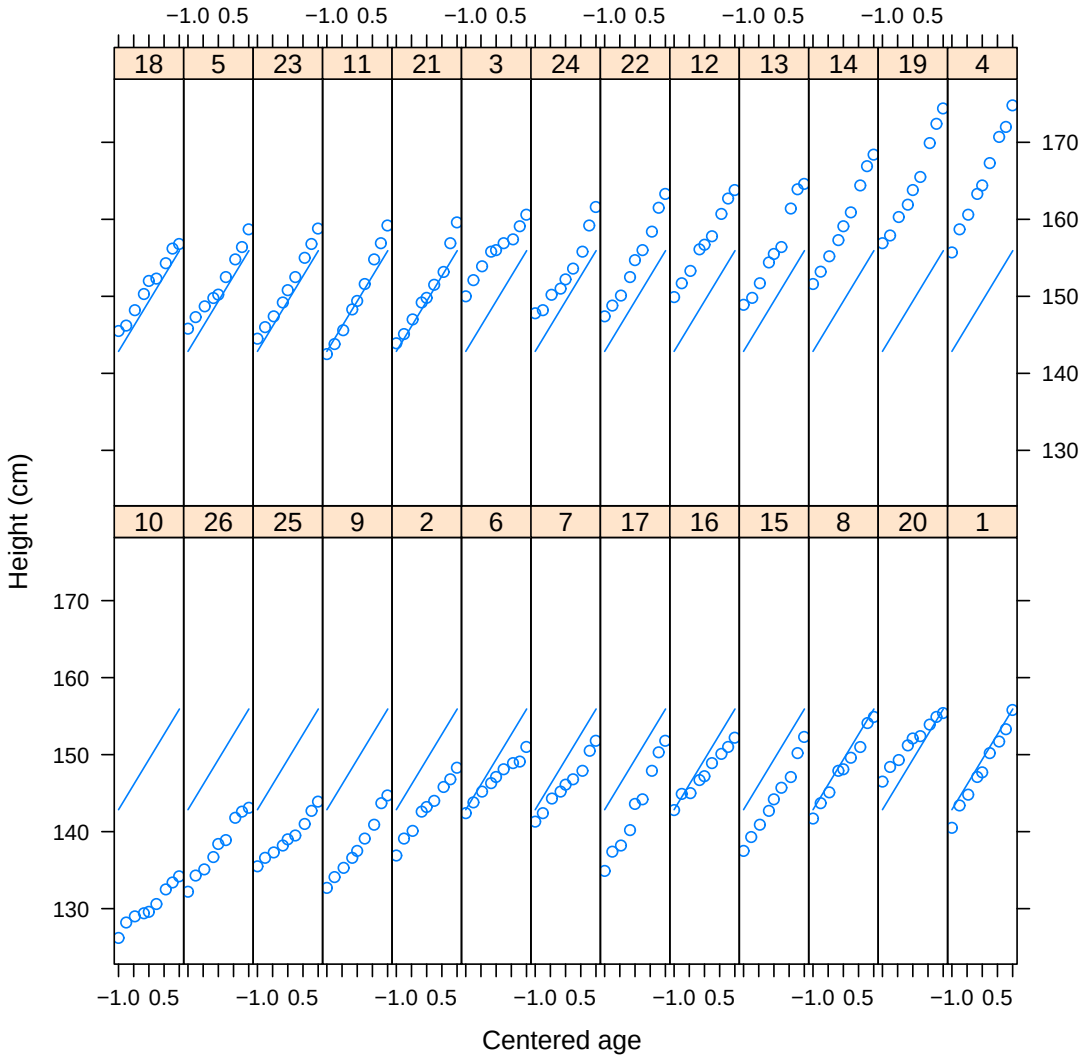

Figure 2: Simple linear regression model fitted to complete dataset

To make our model more realistic, we allow the intercept in Model 1 to vary from subject to subject. Writing H_{ij} for the ith measurement on the jth subject, we have ...

$$H_{ij} = \beta_{0j} + \beta_1 A_{ij} + \epsilon_{ij} \tag{2}$$

Notice that the intercept β_{0j} now has a subscript j, indicating that it will vary from subject to subject.

We now assume that the individual intercepts follow a Normal distribution with variance τ_0. This gives the model...

$$\beta_{0j} = \beta_0 + u_{0j} \tag{3}$$

... where $u_{0j} \sim N(0, \tau_0)$

Model 2 accounts for the variation in the individual measurements on a single subject, while Model 3 accounts for the variation from one subject to another. The combination of these two models gives what is known as a *multilevel model*.

Fitting our multilevel model to the data in Figure 1, we obtain the predictions shown in Figure 3.

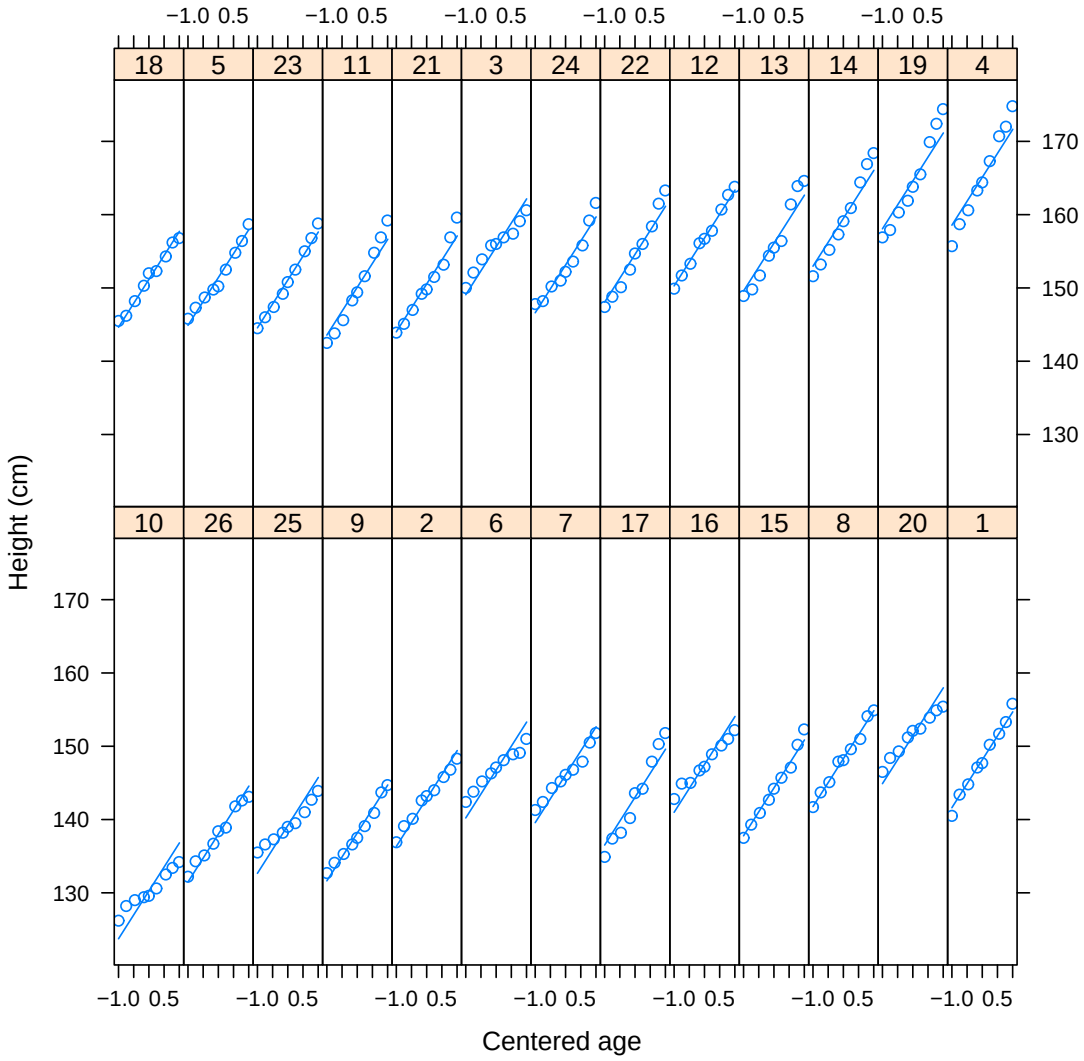

Figure 3: Multilevel model with varying intercept

The fit is much better than for the simple linear regression model, but there is still some evidence of lack of fit. Although we're allowing the intercept to vary from subject to subject, we're using a common slope. As a result, we're overestimating the growth rate of some subjects - e.g. subject 10, and underestimating the growth rate of others - e.g. subjects 4 or 19. Perhaps we can improve our model by allowing both the intercepts and slopes to vary randomly. Figure 4 shows the predictions from a model of this kind.

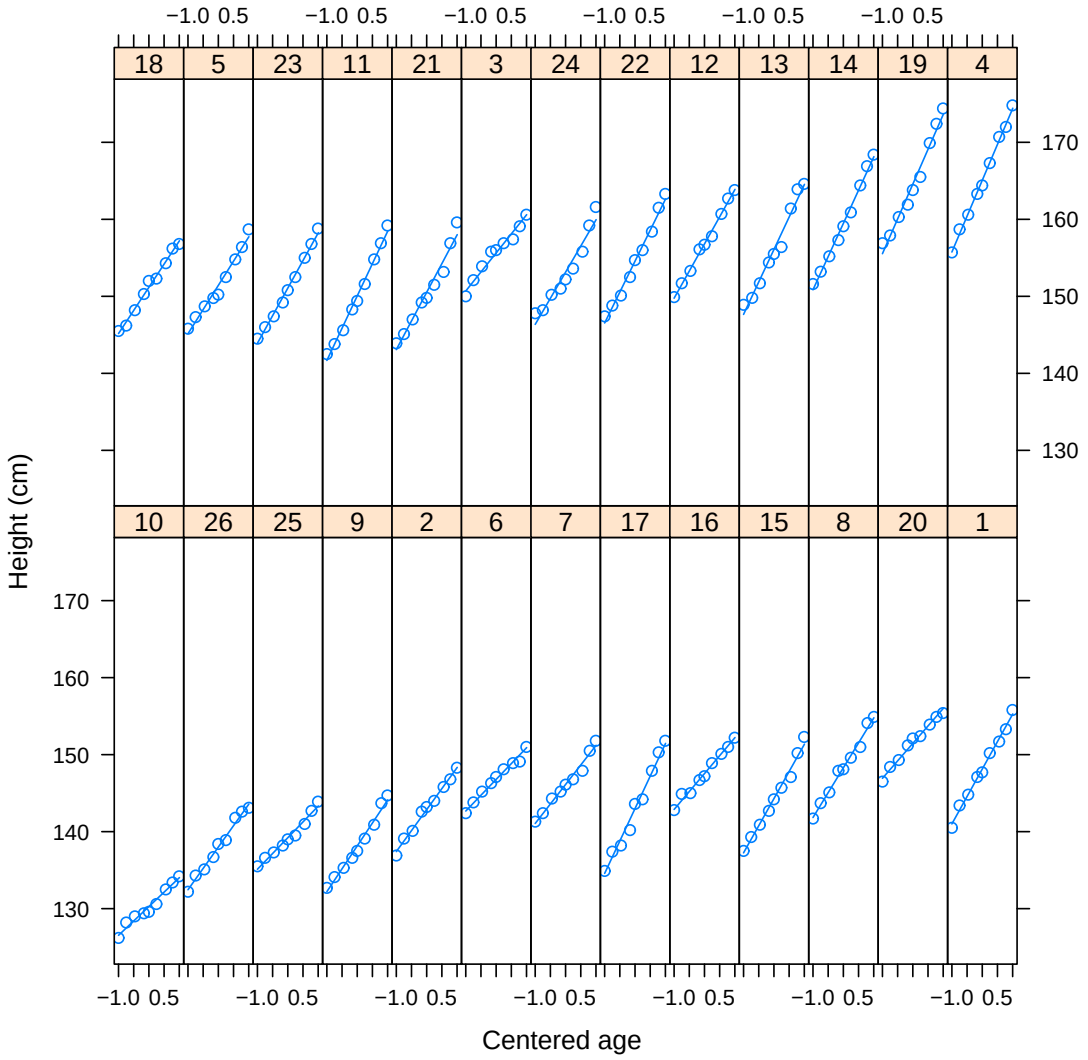

Figure 4: Multilevel model with varying intercept and slope

3 How do multilevel models differ from regression models?

To show the difference between a multilevel model and an ordinary regression model, we return to the model with varying intercepts and substitute equation 3 into equation 2 to give...

$$H_{ij} = (\beta_0 + u_{0j}) + \beta_1 A_{ij} + \epsilon_{ij} = \beta_0 + \beta_1 A_{ij} + u_{0j} + \epsilon_{ij} \tag{4}$$

The feature that distinguishes this model from an ordinary regression model is the presence of *two* random variables - the measurement level random variable ϵ_{ij} and the subject level random variable u_{0j}.

Because multilevel models contain a mix of *fixed* effects and *random* effects, they are

sometimes known as *mixed-effects* models.

4 Benefits of multilevel modelling

In a multilevel model, we use random variables to model the variation between groups. An alternative approach is to use an ordinary regression model, but to include a set of dummy variables to represent the differences between the groups. The multilevel approach offers several advantages.

- We can generalize to a wider population
  - e.g. can say something about the growth curves that we might expect in the population of boys from which our sample was selected
- Fewer parameters are needed
  - With the height data, we only needed one additional parameter - the variance of the $u_{0,j}$ - in order to allow the intercepts to vary from subject to subject. By contrast, the approach via dummy variables would require 25 additional parameters. This reduction in the number of parameters is particularly important with more complex models and a limited amount of data.
- Information can be shared between groups
  - By assuming that the random effects come from a common distribution, a multilevel model can share information between groups. This can improve the precision of predictions for groups that have relatively little data.

5 Extending multilevel modelling

This section indicates some of the ways in which multilevel models can be extended to deal with more complex behaviour.

- Inclusion of predictors at the group level
  - e.g. if our height data included both boys and girls, we could include a term in the group level model to describe the difference in height between boys and girls
- Multiple levels of grouping
  - e.g. pupils within classes within schools
- Multivariate responses
  - e.g. exam results in English *and* Maths at age 16

- Cross-classified data
  - e.g. children in one school may come from different areas and children from the same area may go to different schools
- Usual regression extensions
  - e.g. multiple predictors, categorical predictors and responses, etc

6 Software for multilevel modelling

There is a wide range of software available for fitting multilevel models - for detailed reviews, see the website of the Centre for Multilevel modelling at the University of Bristol.

7 References

For a short introduction to multilevel modelling, see Browne and Rasbash (2004). For a more detailed treatment, see Pinheiro and Bates (2000) or Raudensbush and Bryk (2002). Both these texts contain extensive discussions of the application of multilevel modelling to real problems - the examples in Raudensbush and Bryk are taken from the social and human sciences, while those in Pinheiro and Bates involve areas such as ergonomics, agriculture and medicine.

Another useful source of information is the website of the Centre for Multilevel Modelling (CMM) at Bristol University. This contains extensive training material and reviews of statistical software.

Browne, W. and Rasbash, J. (2004). 'Multilevel Modelling', in Hardy, M. and Bryman, A. (eds.), *Handbook of data analysis*, Sage Publications, pp 459-78.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*, Springer.

Raudensbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models*, Sage Publications.