Chapter 5
Graphs and Equations

Abstract Although we have introduced the conceptual basis for multilevel analysis in earlier chapters, it remains a statistical method; this chapter introduces the statistical principles of MLA. This is done primarily through algebraic notation, and the equations are linked to graphs where appropriate to help with the interpretation. We build up the chapter from a single-level regression analysis to a random intercept model and finally to a random slope model. We introduce the idea of intraclass correlation and provide visual examples of typical patterns of covariance between the intercept and slope residuals. We look at simple extensions to a third level and the use of complex variance functions to account for heteroscedasticity, and finally we draw comparisons between fixed effects and random effects models.

Keywords Multilevel analysis · Single-level regression · Random intercept model · Random slope model · Intraclass correlation · Variance · Covariance

Multilevel analysis is, as we have discussed, a form of regression analysis that is appropriate when the assumption of independence of observations that underlies ordinary regression models does not hold. The reason for this assumption being violated is the influence of the context; Chap. 4 has introduced a variety of contexts that may be important for our analyses and which may extend beyond 'typical' contexts such as neighbourhood, hospital or school to include, for example the individual (for repeated measures or multiple responses) or time (for repeated cross-sections).

We start this chapter with the basic, single-level, linear regression model and show how we can change this into a multilevel model by adding the context. As the chapter progresses, we cover a range of multilevel models and introduce some of the commonly encountered ideas and terminology such as the intraclass correlation coefficient and random slopes. Where possible we link these ideas to graphs as an aid to interpretation.

The chapter works through the random intercept and random slope models based on the example introduced in Chap. 3 concerning an investigation of the relationship between the time spent on exercise each week and certain individual and contextual characteristics. In this example, we have data that were collected in a health

A. H. Leyland, P. P. Groenewegen, *Multilevel Modelling for Public Health and Health Services Research*, https://doi.org/10.1007/978-3-030-34801-4_5

interview survey. The respondents' addresses were geo-coded, and in this manner, the respondents were allocated to neighbourhoods in the study area. We provide the algebraic notation of the regression equations and introduce the basic terminology cumulatively as we progress. For reference, this terminology is summarised in Box 5.3 at the end of this chapter.

Ordinary Least Squares (Single-Level) Regression

Using a single-level regression model, we would regress the dependent variable, the time spent exercising each week, on one or more independent variables ignoring the neighbourhood in which people live, and how this may affect our outcome. Consider a regression including only the respondent's age; the regression equation is

$$y_i = \beta_0 + \beta_1 x_{1i} + e_{0i} \tag{5.1}$$

In this equation, y_i is the dependent variable. Note that for the single-level regression model, we do not pay any attention to the area of residence of each individual and, as such, the dependent variable is uniquely identified by the subscript *i*. β_0 is used to denote the intercept—the number of minutes spent exercising by the reference group: respondents for whom all independent variables take the value 0. (The value 0 may not always be the best choice; in terms of respondent's age, for example we would not be interested in the time spent exercising by respondents who are 0 years old. To overcome this problem, we may choose to centre some of the independent variables such as age, so that the intercept takes on a more meaningful value, such as the time spent exercising by a respondent of average age. See Chap. 11 for an example of this in practice.) x_{1i} is the independent variable, in this case the age of respondent *i*. β_1 indicates the average change in time spent exercising per week associated with a 1 year increase in age. e_{0i} is the residual or error term.

This equation is illustrated graphically in Fig. 5.1. The time spent exercising tends to decrease with increasing age; the extent to which there is a decrease is determined by the slope β_1. The error term e_{0i} is the vertical distance between the regression line and each observation; in other words, it is the difference between the time that we would expect individual *i* to spend on exercise given their age, $\beta_0 + \beta_1 x_{1i}$, and the time that they actually spent on exercise, y_i.

Equation (5.1) is accompanied by an important assumption about the residuals e_{0i} namely that they are identically and independently distributed and can be characterised by a normal distribution with mean 0 and variance σ^2_{e0}.

$$e_{0i} \sim N\left(0, \sigma^2_{e0}\right) \tag{5.2}$$

In this equation, *N* indicates that the residuals are assumed to follow a normal distribution with zero mean and variance σ^2_{e0}.

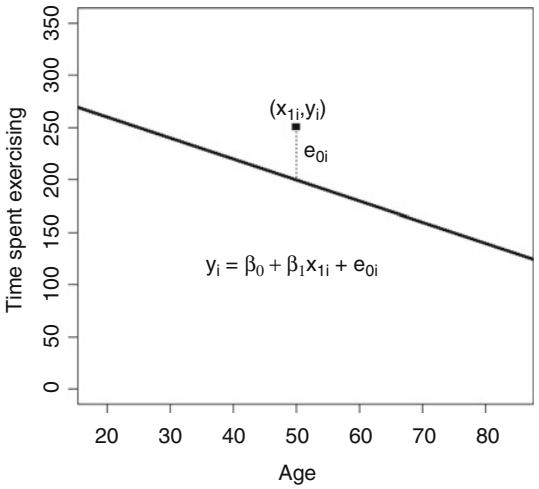

Fig. 5.1 Ordinary least square regression

As we described in Chap. 3, this error distribution is often seen as being nothing more than a nuisance; it is, after all, the part which cannot be explained by our model. But the assumption that the residuals are independent of each other is the one that we are in danger of violating if there is a level missing from our model—neighbourhood in this example. This leads us to the random intercept model.

Random Intercept Model

In a random intercept regression model, we include an effect for each area that impacts on all individuals in that area equally, regardless of their age.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij} \tag{5.3}$$

In this equation, the new terms introduced to Eq. (5.3) over and above those in Eq. (5.1) are as follows. y_{ij} is our dependent or response variable: the outcome for individual i living in neighbourhood j, the number of minutes per week spent exercising. Our survey respondents are numbered from $i = 1, \ldots, N$ and each lives in one neighbourhood $j = 1, \ldots, J$. There are n_j respondents in neighbourhood j so $N = \sum_{j=1}^{J} n_j$. x_{pij} are the independent or explanatory variables, again measured on individual i in neighbourhood j. The subscript p is used simply to distinguish between the different variables; for example x_{1ij} might be the individual's age in years and x_{2ij} a dummy variable indicating the subject's sex (1 = male, 0 = female). x_{pj} are also independent variables, but these are measured at the contextual or neighbourhood level; that is, they take the same value for all individuals living in neighbourhood j. These variables may be directly observed or measured at the neighbourhood level; for example, x_{3j} may be the proportion of the surface area of

neighbourhood j that is characterised as being 'green space'. Alternatively, the contextual variables may represent an aggregation of individual measures; x_{4j} may be the average age of the respondents in neighbourhood j.

β_p is the regression coefficient associated with x_{pij} or x_{pj}. So β_1 would indicate the average change in time spent exercising per week associated with a 1-year increase in age and β_2 would show the average effect of being male on the time spent exercising (relative to that for the baseline category, female, for which $x_{2ij} = 0$). u_{0j} is the estimated effect or residual for area j. This is the difference that we expect to see in the time spent exercising for an individual in neighbourhood j compared to an individual in the average neighbourhood, after taking into account those (individual or neighbourhood) characteristics that have been included in the model. The 0 in the subscript denotes that this is a *random intercept* residual, a departure from the overall intercept β_0 applying equally to everyone in neighbourhood j regardless of individual characteristics. e_{0ij} is the individual-level residual or error term for individual i in neighbourhood j.

Figure 5.2 illustrates this equation graphically. As in Fig. 5.1, the time spent exercising for someone living in an average area is shown as the heavy line, and this relationship is determined by just the person's age x_{1ij}. The part of Eq. (5.3) involving the β coefficients, $\beta_0 + \beta_1 x_{1ij}$, is called the *fixed part* of the model because the coefficients are the same for everybody; the residuals at the different levels, $u_{0j} + e_{0ij}$, are collectively termed the *random part* of the model as these values depend on the individual and neighbourhood. The additional effect for inhabitants of area j, u_{0j}, applies to all inhabitants of the area regardless of age; people in the area illustrated in Fig. 5.2 tend to do more exercise than average. The time we would expect individual i to spend on exercise now depends on their area of residence and

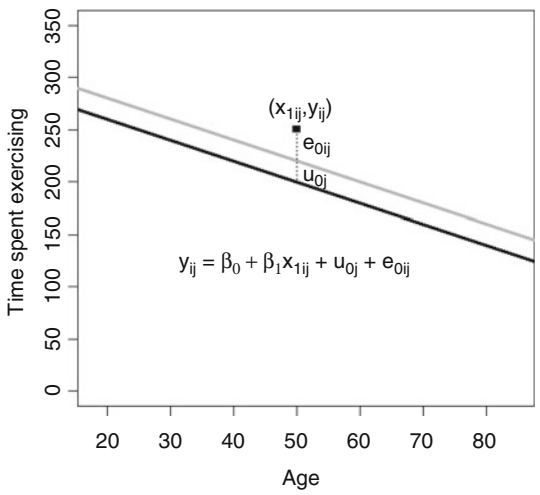

Fig. 5.2 Random intercept model

is given by $\beta_0 + \beta_1 x_{1ij} + u_{0j}$; this is shown in Fig. 5.2 as the grey line. The vertical distance between the two lines, u_{0j}, is constant (i.e., it does not depend on age).

In Fig. 5.2 we can see that the vertical distance from the observed time that person i in area j spends on exercise, y_{ij}, and the average time that someone of this age would spend on exercise, $\beta_0 + \beta_1 x_{1ij}$, is now broken down into a part that is due to the difference between area j and the average, u_{0j}, and a part that is due to the difference between individual i and the average for area j, e_{0ij}. Both the components have their associated distributions and variances:

$$\begin{aligned} u_{0j} &\sim N\left(0, \sigma^2_{u0}\right) \\ e_{0ij} &\sim N\left(0, \sigma^2_{e0}\right) \end{aligned} \tag{5.4}$$

In this equation, σ^2_{u0} is the variance of the neighbourhood-level intercept residuals u_{0j}.

In Eq. (5.3) the fixed part of the model $\beta_0 + \beta_1 x_{1ij}$ does not vary given a person's age x_{1ij}. The total unexplained variation in the outcome (adjusted for age) is therefore equal to the variance of $u_{0j} + e_{0ij}$ or $\sigma^2_{u0} + \sigma^2_{e0}$; that is, some of the variation in time spent exercising is due to differences between neighbourhoods and some is due to the differences between individuals within neighbourhoods. Figure 5.2 shows how the time spent exercising varies with age on average across all areas (black line) and also in area j (grey line). Figure 5.3a shows the relationship for all areas in our sample; each area is shown as a separate line. The variability between areas is then the extent to which these lines are dispersed around the average; if the lines are close together, then there is little variation between neighbourhoods and σ^2_{u0} is small.

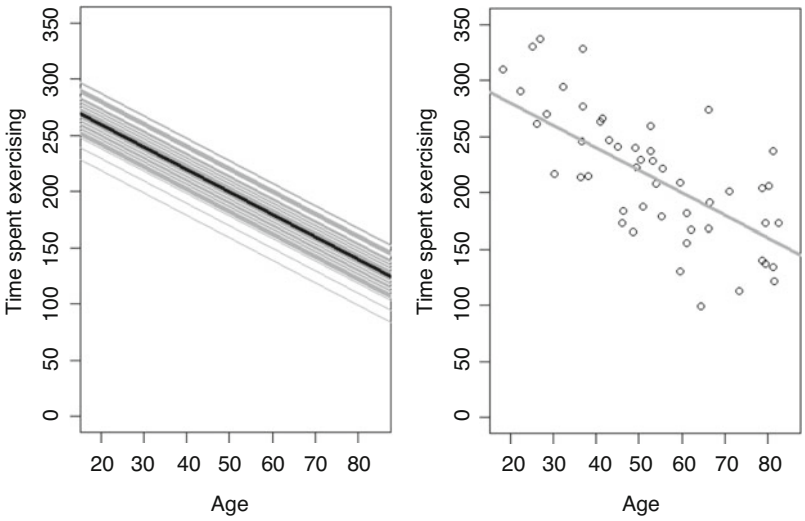

Fig. 5.3 Random intercept model showing (**a**) variation between neighbourhoods and (**b**) variation between individuals within a single neighbourhood

Figure 5.3b shows the variability of the observations made on respondents living in area *j*; these tend to be higher than average (given the individuals' ages) since the area mean is clearly higher than the population mean shown in Fig. 5.3a. However, there is some variability in the tendency to exercise. Some people spend more time exercising than the average for that age in the area whilst others spend less than average—indeed, some spend less than the population average as there is considerable scattering around the average for area *j*. The variability between individuals within areas is then the extent to which the observations are scattered around the average for each area; if the observations are close to the line, then there is little variation within neighbourhoods and σ^2_{e0} is small.

The proportion of the total variance that is due to differences between neighbourhoods is the intraclass correlation coefficient ρ_{I}:

$$\rho_{\mathrm{I}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}} \tag{5.5}$$

ρ_{I} is a measure of the similarity between two people from the same neighbourhood and will take a value between 0 and 1 inclusive. If there were no variation between the area effects then all of the u_{0j} would be equal (to zero) and σ^2_{u0} would be zero meaning that $\rho_{\mathrm{I}} = 0$. If there were no variation within neighbourhoods (following adjustment for age), then the time spent exercising would be determined exactly by age and neighbourhood alone. In this case, σ^2_{e0} would be 0 and so we can see from Eq. (5.5) that $\rho_{\mathrm{I}} = 1$; the exercise times of individuals from the same area would be perfectly correlated. The size of ρ_{I} varies between studies and is very important for power calculations; we return to a discussion of this in Chap. 6. Typically we might expect somewhere around 2–5% of the total variation to arise due to differences between contexts although there are notable exceptions in public health and health services research when this proportion might be higher. Clustering within families or households tends to be quite strong giving large intraclass correlation coefficients; Cardol et al. (2005) found that 18% of the variance in the frequency of medical contacts was attributable to the family, and Sacker et al. (2006) found 13–21% of the variation in poor general health and 20–34% of the variation in limiting illness was attributable to differences between households. For studies in which the data comprise repeated measures on individuals a large proportion of the variability is often at the individual level (which is not the lowest level in a repeated measures design—see Chap. 4). For example, Lipps and Moreau-Gruet (2010) found that over 90% of the total variance in body mass index was at the individual (as opposed to measurement occasion) level in a repeated measures analysis.

The model described by Eqs. (5.3) and (5.4) is the basic random intercept or variance components model. These terms are used interchangeably which might be confusing when reading studies that report multilevel analysis. There is, however, a glossary of the terminology used in MLA (Diez-Roux 2002) which is useful to have at hand when reading papers that use this technique. As with the single-level regression model, there are certain implicit assumptions regarding the residuals.

As well as assuming that the residuals at each level are independently and identically distributed, the model is built on the assumption that the neighbourhood residuals u_{0j} are independent of the individual level residuals e_{0ij} and that they are uncorrelated with all of the independent variables (in this case x_{1ij}). In a multilevel model described by Eqs. (5.3) and (5.4), it is possible that there will be a correlation between the independent variable x_{1ij} and the neighbourhood residuals u_{0j}. This can be avoided by including the group (contextual) mean $x_{2j} = \sum_i x_{1ij}$, so that Eq. (5.3) becomes

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_{0j} + e_{0ij} \tag{5.6}$$

Random Slope Model

From Fig. 5.3a you will note that whilst the intercept—the point at which the lines cross the vertical axis—varies between neighbourhoods, the slope is the same in all areas. The lines are parallel, indicating that a fixed increase in age is associated with the same average decline in time spent exercising in all areas. A random slope model allows the relationship between the independent and dependent variables to differ between contexts; we enable this by including an area effect for the slope (the relationship between time spent on exercise and age) in addition to the area effect for the intercept.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{0ij} \tag{5.7}$$

The new term in this equation is u_{1j}. This is the slope residual for neighbourhood j that is associated with the independent variable x_{1ij}. Just as u_{0j} denotes a departure from the overall intercept β_0, u_{1j} indicates the extent of a departure from the overall slope β_1 in a *random slope* model. In general, there may be a residual u_{pj} associated with any of the independent variables x_{pij} or x_{pj}. However, not every slope will be random and so there will not be slope residuals for every regression coefficient.

The fixed part of this model is, as before, $\beta_0 + \beta_1 x_{1ij}$, and this is shown as the black line in Fig. 5.4. The random part is now given by $u_{0j} + u_{1j} x_{1ij} + e_{0ij}$ which clearly depends on the individual's age x_{1ij}. The grey line in Fig. 5.4 is determined by the fixed part together with both area effects (the intercept residual u_{0j} and the slope residual u_{1j}), i.e. $\beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij}$. For the selected area, there is still a tendency to exercise more than average; the light line in Fig. 5.4 is consistently above the heavy line. But unlike the random intercept model in Fig. 5.2, the distance between the two lines in Fig. 5.4 varies according to the person's age; the increased mean time spent exercising in area j is greater at younger ages than at older ages. This means that the relationship between time spent exercising and age differs between areas. On average, a 1-year increase in age is associated with a change of

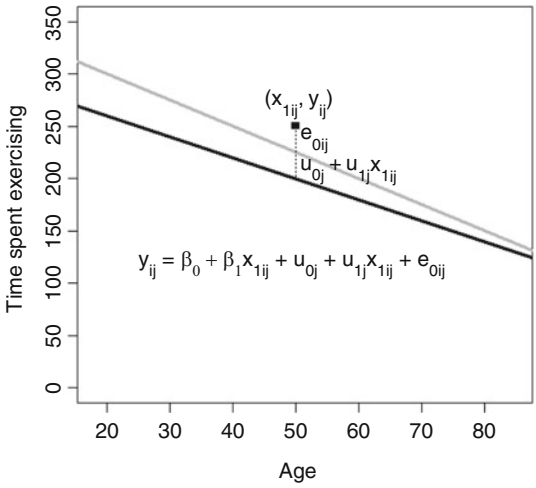

Fig. 5.4 Random slope model

β_1 in the time spent exercising, but in area j, each additional year is associated with a difference of $\beta_1 + u_{1j}$ minutes.

Just as the intercept residuals u_{0j} have an associated variance (σ^2_{u0}), the slope residuals u_{1j} also have a variance (σ^2_{u1}). What is new, however, is the introduction of a covariance (σ_{u01}) between the intercept residual and the slope residual for each area.

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right) \qquad (5.8)$$
$$e_{0ij} \sim N\left(0, \sigma^2_{e0}\right)$$

The covariance is a measure of the extent to which two variables change in the same direction. We can use the covariance between u_{0j} and u_{1j}, along with the two variances, to calculate the correlation between the two:

$$\rho_{u01} = \frac{\sigma_{u01}}{\sqrt{\sigma^2_{u0}\sigma^2_{u1}}} \qquad (5.9)$$

The unexplained variance in Eq. (5.7) is now

$$\text{var}\left(u_{0j} + u_{1j}x_{1ij} + e_{0ij}\right) = \sigma^2_{u0} + x^2_{1ij}\sigma^2_{u1} + 2x_{1ij}\sigma_{u01} + \sigma^2_{e0} \qquad (5.10)$$

The term involving the covariance σ_{u01} takes into account the fact that the intercept and slope residuals, u_{0j} and u_{1j}, are not independent of each other. The covariance matrix in Eq. (5.8)—the variances σ^2_{u0} and σ^2_{u1} and the covariance σ_{u01}—conveys a variety of information about the different relationships between time spent exercising and age for the neighbourhoods in our study. Figure 5.5 shows how

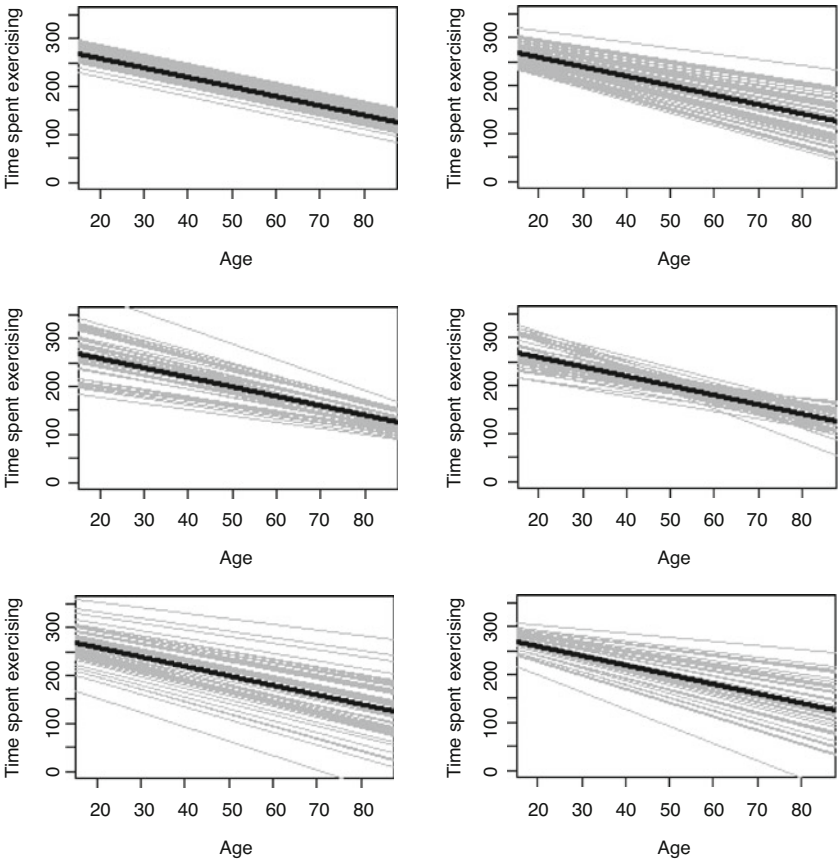

Fig. 5.5 Random slope model with differing covariance matrices showing (**a**) small (or zero) slope variance; (**b**) moderate intercept and slope variance, positive covariance; (**c**) moderate intercept and slope variance, negative covariance; (**d**) moderate intercept and slope variance, small (or zero) covariance; (**e**) large intercept variance, moderate slope variance, positive covariance; and (**f**) moderate intercept variance, large slope variance, positive covariance

various patterns in the covariance matrix can be translated into different graphs illustrating these relationships. The fixed part of the model $\beta_0 + \beta_1 x_{1ij}$ is the same in each graph, and so the black line—denoting the relationship in the average area—does not change. Firstly, Fig. 5.5a shows that if the variance of the slope is very small or zero then we are back to a random intercept model. The lines for the neighbourhoods are parallel to each other since the relationship between exercise and age does not vary between contexts. Figure 5.5b illustrates a moderate slope variance and a positive covariance between the intercept and slope residuals for each area. In general, areas with a large (small) intercept residual u_{0j} will tend to have a large (small) slope residual u_{1j}, meaning that areas with intercepts higher than average will tend to have slopes that are more positive (or less negative) than

average. If the inhabitants of an area tend to do more exercise than average this will usually be the case at all ages, but this benefit is most pronounced at older ages. This leads to the general pattern of lower variability between areas at younger ages and increased variability at older ages. Equation (5.10) shows how the unexplained variance will increase with age if the covariance σ_{u01} is positive. In Fig. 5.5c the covariance between the intercept and slope is negative, meaning that areas with higher intercepts tend to have lower (or more negative) slopes. This leads to a pattern of increased variability between neighbourhoods at young ages and decreased variability at older ages. Figure 5.5d illustrates a case in which the covariance between the intercept and slope residuals is very small or zero (centred around age 50 years: see Box 5.1); in such a case, there is no relationship between the two. Unlike Fig. 5.5b, c, the knowledge that the mean time spent exercising at age 50 years in one particular area is higher than average does not impart any further information about whether the slope will be flatter or steeper than average. The lines for the neighbourhoods cross quite randomly. In Fig. 5.5e, we can see the impact of increasing the intercept variance for the model seen in Fig. 5.5b, and Fig. 5.5f demonstrates the effect of increasing the slope variance again from that seen in Fig. 5.5b. The former tends to increase the average effect or distance from the heavy line (the average area) whilst the latter tends to increase the difference between areas in the strength of the relationship between exercise and age.

The interpretation of the covariance given above is a slight simplification since this actually depends on the centring of the independent variable. This means that the size, and even the sign, of the covariance can change if the independent variable is centred around a different value although neither the data nor the pattern of convergence or divergence of areas will change. See Box 5.1 for an explanation.

Box 5.1 The Effect of Centring on the Covariance

In the equations in this chapter, x_{1ij} is the age of individual i in neighbourhood j, taking values dependent on the sample. In Eq. (5.1), β_0 is the intercept and denotes the time spent on exercise for an individual for whom all covariates are equal to zero; in other words, this is the mean time spent on exercise by a person who is 0 years old. Since this is almost certainly outside the range of our data, we can choose to centre age around another value as an aid to interpretation. To centre around age 50 years, we would replace x_{1ij} by $x^*_{1ij} = x_{1ij} - 50$, so that the random slope model in Eq. (5.7) becomes

$$y_{ij} = \beta^*_0 + \beta_1 x^*_{1ij} + u^*_{0j} + u_{1j} x^*_{1ij} + e_{0ij}$$

The new intercept, β^*_0, now indicates the mean time spent on exercise by a 50-year old. The estimate of the slope, β_1, has not changed and nor have the slope residuals u_{1j}. The u^*_{0j} are the random intercept residuals which now represent area effects for 50-year olds (as opposed to the u_{0j} which were the

(continued)

Box 5.1 (continued)
area effects for those aged 0 years). You can see from the random slope model in Fig. 5.4 that magnitude of the area effect, or the distance from the grey (area-specific) line to the black (population) line, differs by age. So changing the intercept in a random slope model also alters the area-specific intercept residual.

Since the intercept residuals change if we change the intercept, their variance also changes and so does the covariance between the intercept and slope. For a centred model, the level 2 variances and covariances given in Eq. (5.8) become

$$\begin{bmatrix} u^*_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^{*2}_{u0} & \sigma^*_{u01} \\ \sigma^*_{u01} & \sigma^2_{u1} \end{bmatrix} \right)$$

It is straightforward to show that in this example $\sigma^{*2}_{u0} = \sigma^2_{u0} + 100\sigma_{u01} + 2500\sigma^2_{u1}$ and $\sigma^*_{u01} = \sigma_{u01} + 50\sigma^2_{u1}$. The implication of this is that the centring of a variable with a random coefficient will change the covariance and therefore the correlation between the intercept and slope residuals.

The interpretation of random slopes will vary according to the substantive nature of the research but always depends on the nature of the covariance. Damman et al. (2011) give a series of examples of random slope models examining the relationship between healthcare experiences and patient characteristics in a sample of patients drawn from 32 family practices in the Netherlands. They showed a negative covariance between the practice-level intercept and the residual for the patient's age, indicating less variability between practices for older patients; similarly variation decreased with increasing patient health status. Although the relationship between educational level and patient experiences could be seen to vary across practices, there was no correlation between the average experience and the slope across educational level. Finally, a positive correlation between the practice-level intercept and the residual for the patient's ethnicity suggested greater variation in experiences between practices for migrant patients than for those from a Dutch background.

Three-Level Model

The two-level random intercept model described by Eqs. (5.3) and (5.4) can easily be extended to include a third level. Assume that the J neighbourhoods are themselves nested within K towns, and we believe it plausible that people's exercise habits may differ between towns as well as between neighbourhoods within towns. The time spent exercising by individual i living in neighbourhood j of town k, y_{ijk}, then includes an effect or residual for town k, v_{0k}, and is given by

$$y_{ijk} = \beta_0 + \beta_1 x_{1ijk} + v_{0k} + u_{0jk} + e_{0ijk} \tag{5.11}$$

The residuals at the three levels are assumed to be independently normally distributed:

$$\begin{aligned} v_{0k} &\sim N\left(0, \sigma^2_{v0}\right) \\ u_{0jk} &\sim N\left(0, \sigma^2_{u0}\right) \\ e_{0ijk} &\sim N\left(0, \sigma^2_{e0}\right) \end{aligned} \tag{5.12}$$

It is now possible to allow the coefficient of age to vary across towns instead of (or as well as) neighbourhoods by introducing a slope residual v_{1k} in the same manner as we did for the neighbourhood level above.

Heteroscedasticity

In linear multilevel models, as with single-level models, we can allow for heteroscedasticity (also known as complex level 1 variation). The two-level random intercept model described by Eqs. (5.3) and (5.4) makes the assumption that the level 1 variance σ^2_{e0} is constant and independent of the person's age x_{1ij}. It may be that this assumption is too simplistic and inappropriate, and instead of the observations being randomly distributed around the line for each area as in Fig. 5.3b, we find that there is more variability in the amount of exercise undertaken by older respondents. Such a scenario is illustrated in Fig. 5.6.

Heteroscedasticity of this kind can be accommodated by including a further residual term at level 1, e_{1ij}, in a manner analogous to the inclusion of a random

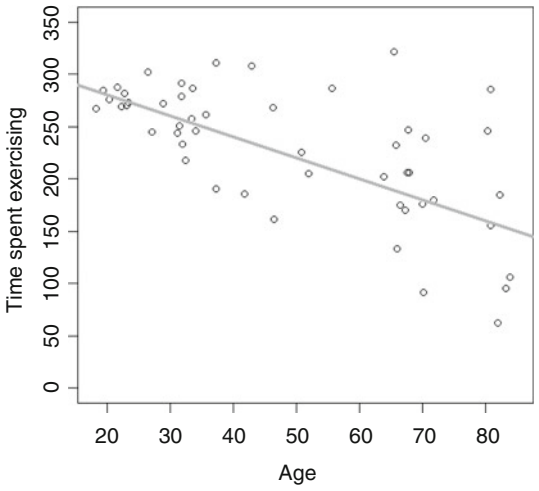

Fig. 5.6 Random intercept model showing variation between individuals within neighbourhoods, with the variance dependent on the respondent's age

slope at level 2: it is only the interpretation that is different. Equations (5.3) and (5.4) now become:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij} + e_{1ij} x_{1ij} \tag{5.12}$$

and

$$\begin{aligned} u_{0j} &\sim N\left(0, \sigma^2_{u0}\right) \\ \begin{bmatrix} e_{0ij} \\ e_{1ij} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{e0} & \sigma_{e01} \\ \sigma_{e01} & \sigma^2_{e1} \end{bmatrix} \right) \end{aligned} \tag{5.13}$$

The unexplained variation in the outcome is now given by the variance of the random part $u_{0j} + e_{0ij} + e_{1ij}x_{1ij}$ which is given by $\sigma^2_{u0} + \sigma^2_{e0} + 2x_{1ij}\sigma_{e01} + x^2_{1ij}\sigma^2_{e1}$. Although the variance between areas is constant, the variance between individuals within areas differs according to the individual's age.

In a single-level regression model, ignoring heteroscedasticity in the data will result in unbiased parameter estimates, but the standard errors associated with these estimates may be incorrect meaning that tests of significance may be misleading. In a multilevel regression model, the failure to model heteroscedasticity that is present in the data may result in the erroneous detection of a random slope (Snijders and Berkhof 2008).

Fixed Effects Model

We introduced the fixed effects model as an alternative to MLA in Chap. 3 and show its algebraic representation here to highlight the differences between the multilevel and fixed effects approaches. Since the fixed effects model introduces a series of $J - 1$ dummy variables to model the effect of the neighbourhoods it is an extension of the single level models described by Eqs. (5.1) and (5.2). We let x_{pi} take the value 1 if individual i lives in neighbourhood p, $p = 2, \ldots, J$, and 0 otherwise. Equation (5.1) then becomes

$$y_i = \beta_0 + \beta_1 x_{1i} + \sum_{p=2}^{J} \beta_p x_{pi} + e_{0i} \tag{5.14}$$

The parameters associated with the dummy variables, β_p, now denote the difference between the mean time spent exercising in neighbourhood p compared to neighbourhood 1 (the baseline). There is only one term in the random part of Eq. (5.14)—e_{0i}—as no assumptions are made about the distribution of the area effects β_p.

When we introduced the fixed effects model in Chap. 3, we mentioned that such models may change the interpretation of the (fixed part) regression parameters. This is because under the fixed effects model, the higher level units are regarded as nuisance parameters and all associated contextual effects are removed from the analysis. However, as described in Chap. 2 when considering the transformation from micro-level to macro-level, the contextual variables available to us include the mean of the characteristics measured at the individual level. The fixed effects model effectively centres all our level 1 independent variables around their mean, so Eq. (5.14) is more appropriately written as

$$y_{ij} = \beta_0 + \beta_1\left(x_{1ij} - \bar{x}_{1j}\right) + \sum_{p=2}^{J} \beta_p x_{pij} + e_{0ij} \tag{5.15}$$

where $\bar{x}_{1j}$ is the average of the x_{1ij} for neighbourhood j. Whilst the parameter estimate β_1 in the multilevel models indicates the association between the time spent exercising and the individual's age, in the fixed effects model β_1 represents the association between the time spent exercising and the extent to which an individual's age differs from the average age of respondents in their neighbourhood. These two effects, and their interpretations, are not necessarily the same (Leyland 2010).

We have tried to ensure that we are internally consistent in terms of the algebraic notation that we use in this book. However, some papers use alternative notations; we describe a common alternative in Box 5.2.

Box 5.2 Alternative Notation Used in MLA

To a large extent the alternative notation used is a substitution of one letter or symbol for another which is trivial if confusing. However, multilevel models are sometimes broken down into separate equations representing distinct parts of the model. This box details the equivalence of the notation that we use in this book to that used by Diez-Roux (2002). We can expand the random slope model given by Eqs. (5.7) and (5.8) to include a contextual variable x_{2j} and the cross-level interaction between the individual and contextual variables $x_{1ij}x_{2j}$:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij}x_{2j} + u_{0j} + u_{1j}x_{1ij} + e_{0ij}$$

The equivalent notation

$$Y_{ij} = \gamma_{00} + \gamma_{10} I_{ij} + \gamma_{01} G_j + \gamma_{11} I_{ij} G_j + U_{0j} + U_{1j} I_{ij} + \varepsilon_{ij}$$

represents a substitution of γ_{00} for β_0, γ_{10} for β_1 and I_{ij} for x_{1ij} etc. and is also sometimes written as

(continued)

Box 5.2 (continued)

$$Y_{ij} = b_{0j} + b_{1j}I_{ij} + \varepsilon_{ij}$$

where

$$b_{0j} = \gamma_{00} + \gamma_{01}G_j + U_{0j}$$
$$b_{1j} = \gamma_{10} + \gamma_{11}G_j + U_{1j}$$

Rankings and Institutional Performance

The higher level residuals in multilevel models are also termed effects because, in the simple case of a random intercept model, the residuals represent the estimated effect of a higher level unit on all of the individuals (level 1 units) contained in that higher level unit. If the levels in a model include an institution such as a care home, school or hospital, then we might like to provide some comparison of institutions to identify those that are performing well or poorly in comparison to their peers—a "league table" of performance. Although the use of performance indicators requires careful consideration and should not be adopted universally (Smith 1995), it is clear that if they are to be used, then their construction should be methodologically sound and that necessitates the use of MLA (Goldstein and Spiegelhalter 1996; Marshall and Spiegelhalter 2001).

In a random intercepts model such as that identified by Eqs. (5.3) and (5.4), the level 2 residual u_{0j} is our estimate of the effect of institution j. As mentioned in Chap. 3, the estimates of the u_{0j} are shrunk towards zero, the mean for all hospitals. The extent of this shrinkage is dependent on the number of observations that we have for any given hospital. The u_{0j} are not known with certainty, hence the need to estimate them. They can typically be plotted together with a measure of uncertainty such as 95% confidence intervals as shown in Fig. 5.7, previously shown as Fig. 2.5; the smaller the confidence interval, the more certain we are about the estimate. Hospital effects in this example comprise the hospital residual u_{0j} added to the mean score for all hospitals, and these are plotted in rank order from the hospital with the lowest mean score (following adjustment for the patient's age, sex, education and physical and mental health) on the left to that with the highest score on the right. Typically there is substantial overlap between the estimates for different hospitals as is the case in Fig. 5.7, meaning that despite having a higher mean score, it is difficult to say with any certainty that one particular hospital is better than a hospital a few positions lower in the rankings.

The production of a measure of institutional performance following adjustment for patient characteristics using a random intercept model can be illustrated by Fig. 5.5a. Although the outcome varies according to the individual's age, the

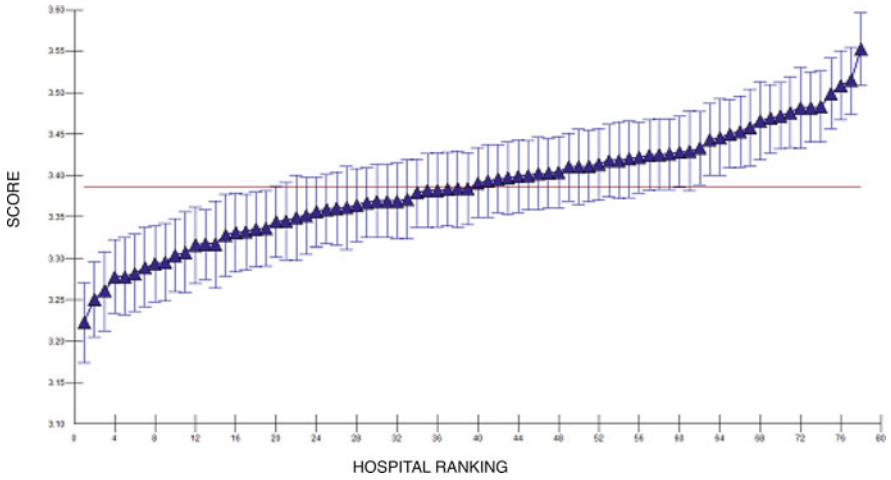

Fig. 5.7 Hospital performance scores (and confidence intervals) for patients' experience of their room and stay (78 hospitals; 22,000 patients). (Source: Sixma et al. 2009)

hospital effect—the distance between the line for any particular hospital and the fixed part of the model (the black line)—is the same for all ages. As a consequence the ranking of the hospitals—the ordering of the lines from lowest to highest—is the same at all ages. With a random slope model, this becomes more complicated; Fig. 5.5d illustrates how the lines in a random slope model may cross each other meaning that the ranking of hospitals will differ according to the patients' age. In the random slopes model defined by Eqs. (5.7) and (5.8), the random part of the model is given by $u_{0j} + u_{1j}x_{1ij}$; this is the composite residual and clearly varies according to the age of the individual x_{1ij}. So in a random slope model, it is unlikely that a single league table would capture all of the differences in rankings, but effects can be estimated (together with confidence intervals) and rankings produced for any given age.

The use of 95% confidence intervals around the residuals in plots such as Fig. 5.7 enables the reader to gauge whether the estimate for any particular unit differs significantly from the effect for the average level 2 unit. Depending on the intended use of such a plot, it may make more sense to adjust the confidence intervals so as to enable comparisons between pairs or sets of units; Goldstein and Healy (1995) describe the mechanics of making such an adjustment.

Conclusion

This chapter has introduced the algebraic notation for the models that are detailed in the rest of the book. The notation system is flexible in that it can readily be extended to include some of the more complex models that were described in Chap. 4. There

are three reasons for needing to understand the algebraic representation of multilevel models. Firstly, it provides a concise means to describe your work in a manner that would enable others to replicate your models. Secondly, it facilitates an understanding of the models used by other researchers when reading literature relevant to your own research. And finally, the algebraic elements introduced in this chapter are the basic building blocks of multilevel regression models constructed using MLwiN, the software used in the practical section of this book (Chaps. 11–13).

Box 5.3 Basic Terminology

This box summarizes the terminology for the various algebraic terms used in the models in this chapter.

y_{ij} is the dependent variable: the outcome for individual i living in neighbourhood j. Individuals are numbered from $i = 1, \ldots, N$ and each lives in one neighbourhood $j = 1, \ldots, J$. There are n_j individuals from neighbourhood j so $N = \sum_{j=1}^{J} n_j$.

x_{pij} are the independent variables, measured on individual i in neighbourhood j. The subscript p is used to distinguish between the variables.

x_{pj} are independent variables, measured at the neighbourhood level; this variable takes the same value for all individuals living in neighbourhood j.

β_0 is used to denote the intercept.

β_p is the regression coefficient associated with x_{pij} or x_{pj}.

u_{0j} is the estimated effect or residual for area j. This is the difference in the outcome for an individual in neighbourhood j compared to an individual in the average neighbourhood, after taking into account those characteristics that have been included in the model. The 0 in the subscript denotes that this is a *random intercept* residual, a departure from the overall intercept β_0 applying equally to everyone in neighbourhood j regardless of individual characteristics.

u_{pj} is the slope residual for neighbourhood j that is associated with the independent variable x_{pij} or x_{pj}. Just as u_{0j} denotes a departure from the overall intercept β_0, u_{pj} indicates the extent of a departure from the overall slope in a *random slope* model.

e_{0ij} is the individual-level residual or error term for individual i in neighbourhood j.

σ^2_{u0} is the variance of the neighbourhood-level intercept residuals u_{0j}.

σ^2_{up} is the variance of the neighbourhood-level slope residuals u_{pj}.

σ_{u0p} is the covariance between the neighbourhood-level intercept residuals u_{0j} and slope residuals u_{pj}.

σ^2_{e0} is the variance of the individual-level errors e_{0ij}.

ρ_I is the intraclass correlation coefficient or the proportion of the total variation in the outcome that is attributable to differences between areas.

References

Cardol M, Groenewegen PP, de Bakker DH, Spreeuwenberg P, van Dijk L, van den Bosch W (2005) Shared help seeking behaviour within families: a retrospective cohort study. Br Med J 330:882–884

Damman OC, de Boer D, Hendriks M, Meuwissen LE, Rademakers J, Delnoij DMJ, Groenewegen PP (2011) Differences between family practices in the associations of patient characteristics with health care experiences. Med Care Res Rev 68:725–739

Diez-Roux AV (2002) A glossary for multilevel analysis. J Epidemiol Community Health 56:588–594

Goldstein H, Healy M (1995) The graphical presentation of a collection of means. J R Stat Soc Ser A 158:175–177

Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc Ser A 159:385–443

Leyland AH (2010) No quick fix: understanding the difference between fixed and random effect models. J Epidemiol Community Health 64:1027–1028

Lipps O, Moreau-Gruet F (2010) Change of individual BMI in Switzerland and the USA: a multilevel model for growth. Int J Public Health 55:299–306

Marshall EC, Spiegelhalter DJ (2001) Institutional performance. In: Leyland AH, Goldstein H (eds) Multilevel modelling of health statistics. Wiley, Chichester

Sacker A, Wiggins RD, Bartley M (2006) Time and place: putting individual health into context. A multilevel analysis of the British household panel survey, 1991-2001. Health Place 12:279–290

Sixma H, Spreeuwenberg P, Zuidgeest M, Rademakers J (2009) [Consumer quality index hospital stay]. NIVEL, Utrecht

Smith P (1995) On the unintended consequences of publishing performance data in the public sector. Int J Public Adm 18:277–310

Snijders TAB, Berkhof J (2008) Diagnostic checks for multilevel models. In: De Leeuw J, Meijer E (eds) Handbook of multilevel analysis. Springer, New York