Chapter 3
What Is Multilevel Modelling?

Abstract In this chapter, we will introduce the basic methodological background to multilevel modelling in verbal form. The underlying graphs and algebra are not covered until Chap. 5. There are two principal reasons for the increasing popularity of multilevel analysis. Firstly, it is more efficient and uses more of the available information than the alternative approaches of distributing contextual information to all individual observations or of aggregating all individual observations to the contextual level. Secondly, multilevel analysis enables the testing of more interesting hypotheses, especially those referring specifically to variation in outcomes or concerning the interactions between characteristics of the context and of individuals. This chapter also covers the idea of what constitutes a level in multilevel research.

Keywords Multilevel analysis · Random intercepts · Fixed effects · Random slopes · Cross-level interaction · Multilevel hypotheses

In public health, we are often interested in discovering what factors are associated with certain outcomes or what the strength of the relationship is between a variable and an outcome. Such relationships are commonly explored using regression analysis, but standard regression analysis makes certain assumptions that are untenable. Most pertinent among these is the assumption that the outcomes are independent of each other for all of the individuals in our study. We have seen from the previous chapter that the behaviour of individuals often cannot be isolated from the macro context in which they operate: the neighbourhood in which people live or the practice in which physicians work, for example. The influence of the context means that outcomes are unlikely to be independent, violating the assumption on which the standard regression model is based. Our solution is to use MLA to take the different levels into account in our analysis.

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_3

Methodological Background

We use multilevel modelling when we are analysing data that are drawn from a number of different levels and when our outcome is measured at the lowest level. Such a situation arises, for example, when we analyse the self-rated health of individuals, and we want to relate this both to individual characteristics, such as age and social class, and to contextual characteristics, such as the population density of the neighbourhood. If we had only one observation for each neighbourhood—that is, if we had sampled and interviewed just one person in each neighbourhood—and sufficient observations in total, then we would just conduct an ordinary single-level regression analysis. Our observations would be independent of each other; although there may be an influence of the neighbourhood context, our observation of this would differ for each individual in our sample as though it were an individual characteristic. Alternatively, if our entire sample were taken from the same neighbourhood, then we would again be able to treat the observations as though they were independent; although there may be a contextual effect, the identical effect would apply to everyone in our sample.

However, the above sampling designs of one person per neighbourhood or of a sample from a single neighbourhood are unusual ones; more commonly, we will have a number of individuals living in each of a number of neighbourhoods. If the place in which people live influences their health, then the observations are no longer independent. Two individuals living in the same neighbourhood have a common context influencing their self-rated health; as a result, some contribution to self-rated health is common for all individuals living in the same neighbourhood that is not shared by those from other neighbourhoods. The ways in which the environmental contexts in which individuals live or work may influence or constrain behaviour were explored in Chap. 2; an example might be that health behaviours are shared within social networks meaning that there is a common influence on self-rated health.

The average level of self-rated health in any particular neighbourhood may be higher or lower than the average for all neighbourhoods, all other factors being equal. Then within that neighbourhood, some individuals will have self-rated health above the neighbourhood average and some below average. So the overall difference between an individual's self-rated health and the population average will be partly attributable to the differences between neighbourhoods and partly due to the differences between individuals within neighbourhoods. When we look at the differences between individuals in our sample, we use the variance as a summary measure of the total variation. The first important feature enabled by multilevel analysis is the ability to split up or partition this variation into that part which is attributable to the neighbourhood and that which is attributable to the individual. The neighbourhood part of the variation consists of the variation of the average self-rated health of each neighbourhood around the overall average. In multilevel analysis, the neighbourhood averages are assumed to be sampled from a distribution of all neighbourhood averages; this is similar to a random effects analysis of variance

(Gelman and Hill 2007). In regression terms we can think of the neighbourhood average as a regression intercept since this then generalises to the introduction of independent or explanatory variables; the fact that these neighbourhood intercepts are assumed to be drawn from a statistical distribution of all possible intercepts gives rise to the term *random intercepts* model.

Earlier we considered two studies in which we would not need MLA. In the first we sampled one person from each of a number of neighbourhoods. In such a situation, we have no variability within neighbourhoods; the average score in each neighbourhood cannot be distinguished from the score of the single person sampled. In the second example, we took our entire sample from a single neighbourhood; this time there is no variability between neighbourhoods, as the population (sample) mean is equal to the mean observed in that neighbourhood. Neither design enables us to distinguish between the levels of individual and area, and so neither is a true multilevel design.

As discussed earlier, the assumption that our observations are independent is violated if our data are hierarchically structured, and we believe that the context may influence the outcomes; the shared context introduces a correlation between two individuals from the same neighbourhood. This has consequences both for the estimation of regression coefficients—measures of the relationships between individual or contextual characteristics and outcomes—and for the standard errors of these estimates (our measures of precision, which determines the extent to which we find a relationship to be statistically significant). Failing to take into account the correlation between individuals within their contexts leads to the phenomenon known as *misestimated precision* (Aitken et al. 1981); ignoring the clustering of individuals within higher level units leads to an overestimation of the effective sample size and hence the tendency to find more relationships significant at a given significance level than the data can actually support.

The *random intercepts* regression model is based on the assumption that, whilst the intercept or average outcome for individuals with a given set of characteristics varies between higher level units, the relationship between the dependent and independent variables is consistent across all contexts. Returning to the example of how self-rated health varies across neighbourhoods, we might find a relationship with income such that those with higher incomes tend to enjoy better health. A linear relationship would suggest that for every unit increase in individual income, we can expect to see a fixed increase in self-rated health. The use of a *random intercepts* model would be based on the assumption that such a relationship between income and self-rated health holds in all neighbourhoods despite health on average being higher or lower in some neighbourhoods. A *random slopes* or *random coefficients* model allows us to relax this assumption and to let the relationship between self-rated health and income vary across contexts; in some neighbourhoods, the health gain associated with a fixed increase in income may be larger than in others. As with the intercepts, the slopes—the relationship between health and income in each neighbourhood—are assumed to come from a distribution of all possible slopes. Moreover, we can examine the relationship between the intercepts and slopes to see whether, for example, the health gain associated with a fixed increase in income is larger or smaller among neighbourhoods in which the average health rating is lower.

Why Use Multilevel Modelling?

We can think of a number of alternatives to multilevel analysis. The most common of these are:

- Aggregate or ecological analysis: ignore the level of the individual and restrict the analysis to the relationship between contexts
- Individual analysis: ignore the effect of context on our estimates of relationships and their associated precision
- Separate individual analyses within each higher level unit
- Individual level analysis with the inclusion of dummy variables to estimate the effect of each higher level unit

As we mentioned in Chap. 2, these alternative approaches may easily lead to inferences at the wrong level, the ecological and atomistic fallacies (Diez-Roux 1998).

Aggregate Analysis

Imagine that we are interested in examining the relationship between the time spent undertaking recreational physical exercise each week and certain individual characteristics (including age, sex, education and income) and environmental characteristics (including area deprivation and the availability of green spaces). The aggregate analysis would involve averaging the time spent exercising by individuals in each neighbourhood and regressing these means on averages of the individual variables (average age, proportion of males, average education and average income) as well as the contextual variables. Such an analysis involves considerable loss of power since the number of observations in our data set is reduced from the total number of individuals to the total number of neighbourhoods in our study. But, more importantly, the analysis may be misleading; the average income in a neighbourhood may reflect opportunities available to everybody in the area (Diez-Roux 1998) and as such may exhibit a different relationship from that seen with individual income. We return to this issue in our discussion of context and composition in Chap. 7 and provide an example of the way in which aggregated individual variables can take on a different meaning in the practical work in Chap. 13.

Individual Analysis

As we have discussed above, conducting the analysis at the individual level when the context is important, and outcomes are therefore correlated, causes problems with misestimated precision (Liang and Zeger 1993). This can be illustrated most easily for (although is not restricted to) *contextual variables*; that is, variables that have been observed, measured or created at the higher level. Whereas in the above example we have measures of education and income for every participant in the study, the contextual variables—area deprivation and the availability of green spaces—are measured at the area level. The number of observations available on each is therefore limited to the number of neighbourhoods in the study and not the number of individuals. Yet in an individual analysis, we would behave as if we had taken a measure of area deprivation for every study participant, resulting in artificially small standard errors and confidence intervals around those regression coefficients. We show the potential effect of even a small degree of clustering on sample size calculations when we consider the importance of variation at different levels in Chap. 6.

Separate Individual Analyses Within Each Higher Level Unit

If the analysis is conducted separately for every high level unit, then this is fine as far as it goes. We can overcome the effects of the clustering of individuals within contexts by making each analysis context-specific. But there are severe limitations to such an analysis. Firstly, we are unable to share relevant information across contexts. So if, for example, the gender effect—the difference between the mean time spent exercising each week for men and women—does not differ significantly between areas, then the separation of the analysis into specific blocks means that we have lost the ability to estimate a single shared regression coefficient. In general we will estimate a complete set of regression coefficients for each neighbourhood. So a regression on four independent variables—plus an average or intercept term—will be undefined without a minimum of five observations in each area. (In practice we would probably want considerably more than five observations if we were to estimate five parameters; a rough guide is to have ten observations per parameter being estimated, meaning that a more realistic minimum might be 50 observations per area.) But secondly, and more importantly, we have lost the ability to estimate contextual effects. Our contextual variables do not vary between individuals within neighbourhoods and so we are unable to estimate directly the effect that area deprivation or the availability of green space has on recreational exercise. A two-stage "slopes-as-outcomes" approach was developed to enable the combination of such separate regression coefficients, and even to permit the introduction of contextual effects to explain variation in regression parameters between context, but such an approach has several notable limitations (Raudenbush and Bryk 1986).

Individual-Level Analysis with Dummy Variables

Our final alternative to fitting a multilevel model is to fit a fixed effect—a dummy or indicator variable—for every higher level unit in our model. This is rather inefficient in that it can require a large number of dummy variables. Fitting a dummy variable to model the intercept in each neighbourhood may not stretch modern computational capability; however, if a dummy variable were required for every household in a study of individuals nested within households, then the large number of single-person households would result in a large proportion of the total available degrees of freedom being used up in a very unparsimonious model. This would effectively remove the characteristics of individuals living in single person households from our model. The equivalent of a *random slopes* model would require a further dummy variable to estimate the regression coefficient for each neighbourhood. But once again the biggest problem with this approach is the inability to estimate the relationship between a contextual variable and the individual outcome. The inclusion of $(n - 1)$ dummy variables to model the intercepts for n neighbourhoods means that there are no remaining degrees of freedom at the neighbourhood level. It is for this reason that these "*fixed effects*" models (as opposed to *random effects* or multilevel models) can only be used to adjust for the potentially confounding influences of contexts on individual-level relationships rather than to explore contextual influences per se. *Fixed effects* models may also change the interpretation of regression parameters in subtle but important ways, particularly regarding the analysis of panel (repeated measures) data (Leyland 2010).

What Is a Multilevel Model?

By now it should be clear that a multilevel model is a form of regression model that is appropriate when the data have some form of a hierarchical structure. We have also covered what a multilevel model is not, including the *fixed effects* model that uses dummy variables to remove the effects of higher level units. But how do multilevel models work? The key is in the distributional assumption made about the higher level units. Rather than estimate a mean for each higher level unit, as is necessary when using a *fixed effects* model, a multilevel model summarises the distribution of the higher level units using a population mean for all contexts and a variance. A single-level regression model already estimates the mean (or intercept), so the additional requirement of a two-level multilevel model is just one parameter—the variance—regardless of the number of higher level units. When we turn a random intercepts model into a random slopes model, rather than including an additional parameter (the dummy variable modelling the slope) for each of $(n - 1)$ neighbourhoods, we need to add just two parameters—the variance of the slopes and the covariance between the intercepts and slopes. This reduction in the number of parameters required means that multilevel models provide a more efficient approach to data analysis.

But how much information is there in a variance? Is this sufficient for our needs? Often we require estimates of the effects or residuals at higher levels in our model; an example would be for models of institutional performance or the "league tables" discussed in Chap. 2. If we are not estimating the effect of each hospital, we can still use multilevel modelling to make inferences about the performance of contexts, such as hospitals. The distributional assumption that we make about the higher level units—usually that they are normally distributed—means that the estimated effect for each unit is shrunk towards the mean for all units. The extent to which the estimated effect for a particular hospital is shrunk towards the overall mean depends on two factors: the extent of clustering in our data and how much information we have about that hospital. The extent of the clustering can be summarised in a simple fashion by the intraclass or intraunit correlation coefficient—the proportion of the total variance that is attributable to the higher level units. Returning to our earlier example, this is the proportion of the variance between individuals in the time spent exercising that is attributable to neighbourhoods. The intraclass correlation coefficient, sometimes referred to as the variance partition coefficient (Goldstein et al. 2002), is also a measure of the correlation in outcomes between two individuals in the same higher level unit, ranging between 0 (no correlation—time spent exercising is completely independent of the neighbourhood of residence) and 1 (perfect correlation—all individuals from the same neighbourhood spend exactly the same time exercising, given their individual characteristics). The estimated effect for each higher level unit is then a weighted average of what the data for that particular unit tell us and the population average; with less information about a given neighbourhood, we have little evidence that the effect is different from the average and hence the greater the shrinkage towards the mean. Small units about which we have little information are said to "borrow strength" from the rest of the sample (Ghosh et al. 1998). Of course the amount of information that we have about each unit is reflected in the (un)certainty around any estimate; confidence intervals will be smaller for neighbourhoods for which we have a lot of information. See for example Fig. 2.6 in Chap. 2.

There are numerous published examples comparing multilevel analyses with alternative methods that illustrate how different the results can be and how the results and conclusions that can be drawn from the studies are dependent on the method of analysis employed. We briefly describe three such studies below.

The first example concerns a training programme in diabetes care for GPs. When the data were analysed at the level of the individual patients, the conclusion was that the training programme had a positive influence on diabetes outcomes. However, because the training programme targeted GPs and not patients and because the patients are nested within the GPs, a multilevel analysis was also performed. In this analysis, the training programme was no longer significant (Renders et al. 2001).

Our second example concerns the impact of an indoor dichlorodiphenyltrichl oroethane (DDT) house-spraying programme, introduced at the village level, on individual malaria parasitaemia in Central Highland Madagascar (Mauny et al. 2004). As well as showing that the standard errors (and hence confidence interval s) of estimates were somewhat larger for the multilevel analysis, the authors showed how the population size of the village appeared to be strongly associated with the presence of parasites when using a conventional logistic regression model, but that this relationship was not significant when a multilevel analysis was conducted.

Finally, Moerbeek et al. (2003) considered the analysis of multicentre intervention studies based on the analysis of data collected on children clustered within classes and schools from the Television School and Family Smoking Prevention and

Cessation Project (TVSFP) (Flay et al. 1988). They showed that not only did ordinary (least squares) and *fixed effects* regression approaches tend to underestimate the standard error of the treatment effect on the post-intervention Tobacco and Health Knowledge Scale (THKS), these two approaches also provided incorrect estimates of the treatment effect.

What Is a Level?

In the first two chapters, we have given a number of examples of contexts that are relevant for people's health and for healthcare utilisation. When dealing with multilevel analysis, these contexts are called levels. We define a level as a *sample* (or a total population if the number is too small to use a sample or if all of the data are available) of contexts; moreover, we may have one or more *characteristics* (or variables) that vary between contexts.

Earlier in this chapter we introduced an example in which we focused on the time spent undertaking recreational exercise. We used information about individuals: the length of time spent exercising each week and information about individual demographic and socio-economic factors that might influence the time spent exercising. We also had information about the context in which these individuals live: the neighbourhood. Now we have two levels: individuals and neighbourhoods. The average of the time spent by individuals within each neighbourhood on recreational exercise varies between neighbourhoods, and a random intercepts model assumes that the neighbourhood means are sampled from some hypothetical distribution of all neighbourhood means. Such an exercise assumes that the higher level consists of units that can be meaningfully sampled. In this case, that would be a sample of neighbourhoods from a population of neighbourhoods. In practice we often work with *all* neighbourhoods rather than a sample; in such a situation, these can still be considered a sample for the generalisability of results. The data for each neighbourhood form a sample of data that could possibly have been collected at different times (if the sample had been drawn and interviews conducted a week earlier or a month later, the results would have differed) and allow us to make inferences about those neighbourhoods and neighbourhoods in general.

To summarise, levels comprise units that can be observed, sampled and analysed. These units have characteristics that can either be directly observed and measured, such as the availability of green spaces in a neighbourhood, or aggregated from individual characteristics, such as average income.

The distinction between a level and its characteristics is important. A characteristic, such as the degree of urbanisation of regions, is not a level. Degree of urbanisation may have a number of values; for example, it may be categorised in six classes from highly urban to sparsely populated countryside. (Some statistical software refers to these classes as levels, but these are clearly quite different from the levels that we are talking about in MLA.) We can sample regions from each of the classes of degree of urbanisation to form a stratified sample, but that does not make degree of urbanisation the level. Categories of urbanisation are not something that we would usually sample. We do, however, sample neighbourhoods or municipalities and then categorise them according to urbanisation, or we may stratify the sampling frame by urbanisation and draw a sample of neighbourhoods from each stratum to ensure that all strata are represented. Urbanisation is a variable, and neighbourhoods are units that, among other things, can be characterised by their degree of urbanisation.

In survey research urbanisation can be used at both the individual and municipality level depending on the sampling design. In health interviews among a random population sample, people are asked questions about health-related behaviour and subjective health. Characteristics of the place where people live may also be requested or recorded. The dataset comprises details about the individuals interviewed and a variable concerning the place where they live. It is possible to study the relationship between degree of urbanisation and, for example, mental health. All units are still at the individual level; there is no sampling of municipalities and the identity of the municipality of residence is not recorded just the *nature* or characteristic of the local area. Such a design might be employed to ensure confidentiality using a random dial telephone survey. Alternatively, the sample design of the same health interview survey could be two-stage such that, firstly, a number of municipalities is sampled, and, within each of the sampled municipalities, a sample of interviewees is drawn. The dataset now contains individual data and the identity of the municipality. Characteristics of the municipality can be added from other sources or constructed by aggregating individual variables. The result is a database with sampled units at two levels. (Multistage sampling designs are covered, along with other multilevel data structures, in Chap. 4.)

In survey practice, a simple random sample is often not considered for pragmatic reasons—consider the costs of conducting face-to-face interviews with people dispersed over a large area, such as a country. In such circumstances a staged sample is used. To take this data structure into account, often simple adjustments are made to the standard errors of parameter estimates. With the diffusion of MLA in health-related research, there are now tools enabling us to treat a multistage sample in an appropriate way, and it has become more common to theorise about the way context affects people's health, health-related behaviour, and health service utilisation.

As long as we only see the pragmatic reason of not having to send interviewers to a large number of different places as the rationale for using a two-stage sample design, the higher level in the data structure is just a nuisance. It is important to take the two-stage nature of the sample into account in statistical analysis, because the outcomes for individuals clustered within the same higher level sampling unit may not be independent. However, if we think of the higher level units as a context for human behaviour, they become interesting in themselves.

In intervention studies, to use another example, the intervention can be made at the individual level or at a higher level related to the provider of the intervention, such as a physician, health centre or community. If the intervention is a new drug, and patients are recruited from one site or the administration of the drug is strictly controlled and independent of where the patients get it, we again have a traditional single level analysis. One of the variables, the marker of the intervention, is whether the patients were given the new drug or a placebo. More complicated interventions often require healthcare providers to follow a protocol when treating eligible patients after randomisation. In this case the sampling design might be that physicians or centres are sampled and then patients are recruited among the eligible population that visit these physicians or centres. In such a case there might be differences in the way the intervention is administered, and it is important to take this into account. Often researchers are only interested in the effect of the intervention, in which case they tend to see the higher level as no more than a nuisance. For example, in a discussion of the advantages of MLA over single-level regression when analysing the relationship between patients' age and blood cholesterol levels, Twisk (2006) states ". . . the medical doctor variable was only added to the regression analysis to be corrected for, and there is no real interest in the different cholesterol values for each of the separate doctors" (p. 9).

How Many Units Do We Need at Each Level?

This question is usually more pressing for the higher level units than for the lower level units. Starting with the number of higher level units we need, we can say that it is not an easy question to answer, and there are no clear rules to follow. We will only give a number of considerations.

First of all, the number needs to be sufficient to estimate a mean and a variance. So the question is: with what number of units would we be confident that we can do that? With somewhere around ten higher level units, it would make sense to do so. With a smaller number it is perhaps better to do a single-level analysis and include dummy variables for the higher level units (a fixed effects model). The accuracy of different parameter estimates from a multilevel model, together with their standard errors, may be dependent on the sample size. Maas and Hox (2005) showed that in general estimates were unbiased in two-level linear multilevel models if there were sufficient (at least 50) higher level units. With fewer higher level units, the only estimate that was affected was the standard error of the high level variance.

Secondly, the research question can impact on the number of higher level units needed. If the research question or hypothesis is about the effect of characteristics of higher level units, such as hospitals, then we need enough hospitals to estimate the effect of the hospital characteristics or test the hypothesis. As a rule of thumb you need an additional ten higher level units for each independent variable at this level that you want to include in the analysis. So if you want to test a specific hypothesis and take into account a few confounders at the higher level, the number of higher level units needed quickly increases.

A related consideration has to with the power available to answer specific research questions (discussed further in Chap. 6). The smaller the number of higher level units, the more difficult it is to find an effect of a given size of a characteristic of the higher level units. If you do not want to be too quick to reject a hypothesis—after all, the hypothesis may be true even if you do not find a significant effect of the variable in question—then one option is to use a different threshold when testing the coefficient of a higher level variable (such as $p < 0.10$ instead of the more common $p < 0.05$).

Cost is often an important factor when making decisions about the number of higher level units to be sampled, especially when data collection for each extra higher level unit is very expensive or burdensome. Snijders (2001) shows how costs may be taken into account when calculating the sample size for a multilevel study.

A final consideration is related to the nature of the higher level units. Sometimes only a certain number of higher level units exist. There are only (currently) 28 European Union Member States, 12 provinces in the Netherlands and 14 health boards in Scotland. So if one of these units is relevant for our research, we are restricted in terms of the numbers available.

In general the number of units within each higher level unit is less of a problem. Even with small numbers of lower level units within each higher level unit, we can estimate a mean and a variance. An example where we have small numbers within higher level units is when we study individuals within households (see e.g. Cardol et al. 2005). There are some situations where it is important. An example is when we want to make league tables to inform patients about quality of care in different hospitals. In this case it is important to have enough observations in each hospital to be able to show significant differences between hospitals; our interest is in estimating the hospital effects, and there have to be enough observations in each hospital to estimate these effects reliably. Another example is when we want to construct new independent variables on the basis of individual observations. This is the case in the field of ecometrics (discussed in Chap. 8) where we might want to say something about safety in neighbourhoods on the basis of survey questions answered by individuals and use that as a neighbourhood characteristic in an analysis of the relation between neighbourhood safety and health. In this case the number of individuals is important to reach a satisfactory reliability of the construct "neighbourhood safety". However, in general, if we have a choice, it will be better to increase the number of higher level units than the numbers within the higher level units.

Hypotheses That Can Be Tested with Multilevel Analysis

As we argued in Chaps. 1 and 2, higher level units are important because they define the action space of individuals. Many problems in public health and health services research are related to people's behaviour; people behave within the social and institutional context of, for example, their community or workplace. This context influences the resources and the range of options (opportunities and constraints) that actors have (Groenewegen 1997). The question "Which levels are relevant?" is answered by analysing the research problem and asking: "What kind of opportunities and constraints determine people's behaviour, and in which units are these opportunities and constraints patterned?" The answers to these questions provide us with hypotheses, and we can now examine the kind of hypotheses that we can test using multilevel analysis.

There is a two-sided relationship between the theories that you want to test and the methodology to do so. Researchers usually do not formulate hypotheses that they are unable to test. If important hypotheses come up that cannot be tested with the standard statistical techniques available at the time, then attempts will be made to develop new techniques. As soon as new statistical techniques are disseminated, new hypotheses develop. This general observation also applies to MLA and the hypotheses that can be tested with it.

MLA makes it possible to test different kinds of hypotheses (Leyland and Groenewegen 2003):

- Hypotheses about variation.
- Hypotheses about the relationship between an outcome variable and individual-level independent variables.
- Hypotheses about the relationship between an outcome variable and higher level (contextual) independent variables.
- Hypotheses about cross-level interactions.

Hypotheses About Variation

The first step in MLA is to consider the variation in an outcome and to split this variation into that part that is attributable to differences between individuals and the part attributable to differences between their contexts. The statistical aspects of this will be introduced in Chap. 5. At present, it is sufficient to know that we can analyse how much of the total variance in our outcome variable is determined by the individual level (e.g. patients) and how much by a higher level, such as doctors or hospitals. In this manner we can get a sense as to how *important* each level is. In MLA we stop seeing variance only as a nuisance parameter that describes uncertainty, but we can also focus on the information that it represents (Merlo 2011).

We can therefore also develop hypotheses about where to expect more variation: at the individual level or at the higher level (Merlo et al. 2005). In many practical applications, the majority of the variation will be at the individual level. If we analyse treatment decisions by physicians, it is reasonable to expect there will be substantially more variability between patients than between doctors. Physicians take into account the situation of individual patients and apply their knowledge and skills according to each patient's circumstances. However, if the patient's situation does not strongly influence the physician's course of action, possibly because there is considerable disagreement between physicians as to the relative value of alternative treatments, more variability might be associated with the physicians. So receipt of treatment A rather than treatment B might be more strongly influenced by the physician consulted than by individual patient characteristics or circumstances.

There are other situations in which we might expect more variation to be at the higher level. This is for instance the case with repeated measures data (this and other data structures will be detailed in Chap. 4). When we analyse repeated measures made on the same individuals (the measures are then the lower level units and the individuals the higher level units), most of the variations will tend to be located at the higher level of the individuals themselves. Think, for example, of repeated measures of a subject's weight; there is likely to be more variability between people than between the measures made at different times on the same individual.

We might also be interested in patients treated by physicians who work together in group practices or hospitals. We now have three levels in our model: the patients, the physicians and the practices in which they work or, alternatively, the patients, hospital departments and hospitals. In this case we can develop hypotheses about the partitioning of variation between physicians and their practices or between hospital departments and the hospitals in which they are situated.

De Jong et al. (2006) considered how the hospital in which physicians worked could influence decisions regarding the length of stay of patients treated. Using data relating to patient discharges from all hospitals in New York State for different medical and surgical diagnostic-related groups (DRGs), they developed and tested a hypothesis based on variation. Believing that physicians would adapt to their local operating circumstances, they hypothesised that there would be more variation in length of stay between hospitals than between physicians working in the same hospital. The variation between individual patients, although substantially larger than the variation between physicians or between hospitals, was not of primary concern for this hypothesis.

In a more exploratory analysis, with no prior hypothesis, it is still important to analyse how variation is distributed between levels. This might provide clues as to what mechanisms could potentially explain variation (Merlo et al. 2009). The extent to which variation is distributed over different levels is also highly relevant when it comes to the development of interventions to influence a certain outcome. Think, for example, about patients' evaluation of their hospital stay. These patient evaluations may be influenced by the attending consultant, by the department where the patients were treated and by the hospital as a whole. Some aspects of the evaluation by patients may relate to the consultant level, such as the patients' judgement as to whether they had received sufficient information from their doctor, whilst other aspects, such as the quality of meals, will be related to the hospital rather than the consultant or department. The extent to which variation is distributed over different levels will give an indication as to the starting point for policies designed to improve patient satisfaction with their hospital stays (Hekkert et al. 2009). Zegers et al. (2011) analysed the occurrence of adverse events in hospitalised patients. From the partitioning of the variance between hospitals and hospital departments, they concluded that interventions to reduce adverse events should not only target hospitals as a whole, but also hospital departments.

Sundquist et al. (2011) studied how individual physical activity was related to objective measures of the built environment among a sample in Sweden. Realising the potential importance of neighbourhood as an influence on individual activity levels, given that neighbourhood is a relevant context for physical activity and that it is an environment that might be amenable to intervention, one of the stated aims of the study was to determine the proportion of the variability in moderate-to-vigorous physical activity that was attributable to neighbourhoods. Finding a rather small proportion of the total variation attributable to neighbourhoods, the authors suggested that the role of urban redevelopment in improving activity levels may be limited.

Apart from splitting the variation in an outcome between the different levels in our model, we can also develop hypotheses about differences in variation between groups. Variation across groups is usually seen by researchers as a nasty statistical problem that is best avoided as opposed to a source of hypotheses (Stinchcombe 2005). In their study on the impact of physician behaviour on patient length of stay, de Jong et al. (2006) reasoned that greater dependencies meant that there would be less variability among physicians who practiced in just one hospital (compared to those working in two or more hospitals). They therefore hypothesised that the variation between physicians (within hospitals) would decrease as the proportion of physicians practicing in just one hospital increased; that is, that there would be more variability within those hospitals in which a larger proportion of doctors worked in more than one hospital.

Ohlsson and Merlo (2007) evaluated the effect of the natural experiment of introducing a decentralised drug budget in Scania county, Sweden, using a before and after design. Believing that the increased economic responsibility given to those responsible for prescriptions would lead to efficient drug prescription, they hypothesised that not only would the prescription of recommended statins increase but also that the variation between healthcare centres and healthcare areas would decrease following budget decentralisation.

In a study of regional inequalities in mortality, Leyland (2004) found that the variance between the mortality rates of districts in Great Britain differed between the 11 regions and tended to increase over time, although the increases were not uniform. These variances were used as a measure of inequality within regions and were considered quite separately from the mean mortality rate for each region.

Although there may not be specific hypotheses concerning differences in variability between subgroups, it should be appreciated that *not* testing for differences in the variance is equivalent to assuming that the variance is the same for all subgroups but failing to test this assumption.

The emphasis on variation is a typical feature of MLA. If you are used to analysing your data at a single level with regression analysis, you probably will not consider differences in the variance between subgroups in your data. Ordinary regression analysis only predicts the means and not the distribution (Stinchcombe 2005). The coefficient of determination (R^2) is used to see how much variation is explained by a set of independent variables, but how much variation there was to begin with is usually not discussed. If you usually use analysis of variance, you might be more aware of differences in variation between groups. When you start using MLA, thinking about variation is an important first step. We return to the subject of variation in more detail in Chap. 6.

Individual-Level Hypotheses

In the case of individual-level hypotheses, a relationship is hypothesised between two variables at the same, lower, level. An example would be the relationship between the educational level of a patient and the amount of negotiating the patient initiates in a consultation with the GP. Why would we use MLA in a case like this? Basically because we know that the relationship cannot be adequately estimated without taking the structure of the data into account. We know that there are numerous other influences on what happens in a consultation, some of which are related to the individual patients and some to the GPs. In that sense the hypothesis about the relationship between educational achievement and initiating negotiations is incomplete, and we cannot simply assume that all other influences are the same (or that they only lead to random variation at the individual level).

Apart from the specific relationship between two variables at the lower level, we can also test the hypothesis that only individual characteristics are responsible for differences in outcomes between contexts such as health differences between communities. If individual characteristics related to health cluster in some communities, one might mistake this for differences produced by community characteristics or circumstances. For example, some communities may have poorer health outcomes but at the same time have older populations. MLA makes it possible to distinguish these so-called *compositional* effects from real *contextual* or area effects. This issue will be dealt with in more detail in Chap. 7. One could of course pose the question as to why people with certain characteristics should cluster together as opposed to being randomly distributed throughout areas. The identification of compositional effects therefore does not solve the problem of the importance of individual choice versus material conditions.

Context Hypotheses

In health services and public health research, as opposed to clinical research, we tend to be more interested in hypotheses relating the context to the outcome when applying MLA. We can distinguish between two kinds of contextual variables: those that are aggregated on the basis of individual characteristics at the lower level, such as the average level of education of the members of a group, and those that are only defined as characteristics of the higher level units. An example of the latter would be the number of years that a group has been in existence. This cannot be deduced from the characteristics of the individuals, but can only be observed for the group as a whole. Context hypotheses can refer to either kind of variable. The interpretation, of course, depends on the researcher's substantive theory. We will only give some possible interpretations here, to emphasise the importance of thinking in terms of possible mechanisms underlying a relationship in order to form hypotheses. We make no pretence that these are the only plausible interpretations.

Aggregated Individual-Level Characteristics

In this case, the higher level variable is constructed by aggregating an independent variable from the lower level to the higher level. (We come back to the way we can construct aggregated variables within MLA in Chap. 8.) There are numerous examples and associated interpretations. We will briefly discuss three.

The first example concerns the number diabetics in a GP's practice and how this number—obtained from counting all diabetics within the practice—might influence the regulation of individual patients. The hypothesis could be that the more diabetics there are in a practice, the greater the chances are that an individual diabetic is more poorly regulated. In this case the mechanism would be *competition*: all diabetics in a practice compete for the scarce and finite resource that is the GP's time and, in so doing, they have to divide the GP's time between them. The consequence is that, as the number of diabetics increases, each of them has less time with the GP and so all of them will be worse off.

The second example is substantively the same, but this time the hypothesis is framed the other way around: the more diabetics there are in a practice, the greater the chances are that an individual diabetic is better regulated. In this case the interpretation could be that a GP with more diabetics on their books is more attentive or more *experienced* in the treatment of diabetics and individual patients within that practice have better results as a consequence.

The aggregation of individual characteristics to a higher level may result in different kinds of variables; we could construct a count of the numbers of subjects having a certain characteristic, as in the previous two examples, the average value of a variable such as age, the proportion of subjects that have a particular attribute or trait (such as smoking), or an aspect of the distribution of a variable. The third example addresses this last possibility. There is a large (and much debated) research literature about income distribution and mortality rates. Henriksson et al. (2010) considered the effect of municipal level income inequality on the incidence of AMI in Sweden, adjusting for individual- and parish-level socio-economic characteristics. Income inequality was measured using the Gini coefficient, a statistical measure of dispersion, and the authors hypothesised that increasing municipality-level income inequality would be associated with elevated risk of AMI.

Higher Level Characteristics

In this case direct observations or measurements are made on the higher level units. These higher level characteristics can be indicators of the same processes that are implicated in the examples using aggregated variables. Competition for a GP's time could also be measured using the booking intervals in office consultations; experience in treating diabetics could be measured directly by testing the knowledge or skills of GPs involved in the study.

The number of higher level units may not be very large; as a rule there will be fewer higher level than lower level units. This may make it feasible to use observation or other more qualitative methods, such as the content analysis of documents as a means of constructing higher level characteristics. For example, if we study the effects of characteristics of urban neighbourhoods on the health behaviours of the people in these neighbourhoods, we can go out into the neighbourhoods and observe, for example, aspects of disorderliness. This is feasible with (perhaps) 20 neighbourhoods; however, it would be very costly to collect information on health behaviour through observation of (perhaps) 50 individuals in each neighbourhood. This means that MLA provides opportunities to combine quantitative and more qualitative approaches. A quantitative survey of patients at the individual level, where we usually deal with large numbers, can be combined with qualitative measures at the higher level. It may also be the case that geocoding provides a simple means to link the availability of structures derived from publicly available lists to specified areas (Macintyre et al. 2008).

The big advantage of MLA is that, if contextual information is available, MLA enables the testing of hypotheses about the relationship between contextual characteristics and individual outcomes, whilst simultaneously taking individual influences on health into account. This provides better estimates of the relationship between context and health. This means that, for example, we can analyse the effect of community wealth on population health, taking individual income into account.

Cross-Level Interactions

The fourth type of hypothesis that can be tested using MLA is that relating to cross-level interactions. These are combinations of (or interactions between) variables at different levels. It is the combination of a particular characteristic of the higher level with a particular individual level variable that is hypothesised to have a specific effect on the dependent variable of interest. Below we consider a couple of examples.

In another study of the effect of income inequality on health, Henriksson et al. (2007) hypothesised that manual workers were at higher risk of death than non-manual workers when living in areas of high-income inequality, arguing that such an effect might be supported by both psychosocial and neomaterial explanations. With data on individuals nested within the municipalities of residence, and following adjustment for both individual occupational social class and area income inequality, testing this hypothesis then equated to testing the significance of the interaction between the individual and contextual variables.

Finch et al. (2010) explored whether the relationship between health—measured using allostatic load, a measure of cumulative physiologic stress—and neighbourhood advantage or disadvantage varied according to an individual's educational status. Their hypothesis—that the relationship between context and the individual outcome would vary depending on individual characteristics—again amounted to a test of the significance of the cross-level interaction between a neighbourhood-level education index of concentration at the extremes (ICE) and individual socioeconomic status (operationalised using educational status).

The ability to analyse cross-level interactions is a major advantage of MLA that follows on from the ability to incorporate both individual and contextual independent variables in an analysis. In our thinking and theorising about health and healthcare, the relationships between context, individual characteristics and outcomes are of central importance. MLA affords the opportunity to test our ideas about these relationships.

Conclusion

In this chapter we have covered the basic concepts of multilevel modelling and have explained its potential and application in non-statistical terms. We have also covered the rationale for MLA and an explanation of what it is and how it differs from other regression approaches. We will return to the important subjects of variance and hypothesis testing at later stages in this book; for the moment it is important that you are aware that variance is not just a nuisance (the unexplained part of a model) and that, whether you are interested in formal hypothesis testing or concerned only with exploratory analysis, variances and contexts add new dimensions to research based solely on individual variables.

References

Aitken M, Anderson D, Hinde J (1981) Statistical modelling of data on teaching styles. J R Stat Soc Ser A 144:419–461

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

de Jong JD, Westert GP, Lagoe R, Groenewegen PP (2006) Variation in hospital length of stay: do physicians adapt their length of stay decisions to what is usual in the hospital where they work? Health Serv Res 41:374–394

Diez-Roux AV (1998) Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health 88:216–222

Finch BK, Do DP, Heron M, Bird C, Seeman T, Lurie N (2010) Neighborhood effects on health: concentrated advantage and disadvantage. Health Place 16:1058–1060

Flay BR, Brannon BR, Johnson CA, Hansen WB, Ulene AL, Whitney-Saltiel DA, Gleason LR, Sussman S, Gavin MD, Glowacz KM, Sobol DF, Spiegel DC (1988) The television school and family smoking prevention and cessation project. I. Theoretical basis and program development. Prev Med 17:585–607

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Ghosh M, Natarajan K, Stroud TWF, Carlin BP (1998) Generalized linear models for small-area estimation. J Am Stat Assoc 93:273–282

Goldstein H, Browne W, Rasbash J (2002) Partitioning variation in multilevel models. Underst Stat 1:223–231

Groenewegen PP (1997) Dealing with micro–macro relations: a heuristic approach with examples from health services research. In: Westert GP, Verhoeff RN (eds) Places and people: multilevel modelling in geographical research. Nederlandse Geografische Studies, Utrecht

Hekkert KD, Cihangir S, Kleefstra SM, Van den Berg B, Kool RB (2009) Patient satisfaction revisited: a multilevel analysis. Soc Sci Med 69:68–75

Henriksson G, Allebeck P, Ringbäck Weitoft G, Thelle D (2007) Are manual workers at higher risk of death than non-manual employees when living in Swedish municipalities with higher income inequality? Eur J Pub Health 17:139–144

Henriksson G, Ringbäck Weitoft G, Allebeck P (2010) Associations between income inequality at municipality level and health depend on context—a multilevel analysis on myocardial infarction in Sweden. Soc Sci Med 71:1141–1149

Leyland AH (2004) Increasing inequalities in premature mortality in Great Britain. J Epidemiol Community Health 58:296–302

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

Leyland AH, Groenewegen PP (2003) Multilevel modelling and public health policy. Scand J Public Health 31:267–274

Liang K-Y, Zeger S (1993) Regression analysis for correlated data. Annu Rev Public Health 14:43–68

Maas CJM, Hox JJ (2005) Sufficient sample sizes for multilevel modeling. Methodology 1:86–92

Macintyre S, Macdonald L, Ellaway A (2008) Do poorer people have poorer access to local resources and facilities? The distribution of local resources by area deprivation in Glasgow, Scotland. Soc Sci Med 67:900–914

Mauny F, Viel JF, Handschumacher P, Sellin B (2004) Multilevel modelling and malaria: a new method for an old disease. Int J Epidemiol 33:1337–1344

Merlo J (2011) Contextual influences on the individual life course: building a research framework for social epidemiology. Psychosoc Interv 20:109–118

Merlo J, Chaix B, Yang M, Lynch J, Råstam L (2005) A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. J Epidemiol Community Health 59:443–449

Merlo J, Ohlsson H, Lynch KF, Chaix B, Subramanian SV (2009) Individual and collective bodies: using measures of variance and association in contextual epidemiology. J Epidemiol Community Health 63:1043–1048

Moerbeek M, van Breukelen GJP, Berger MPF (2003) A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. J Clin Epidemiol 56:341–350

Ohlsson H, Merlo J (2007) Understanding the effects of a decentralized budget on physicians' compliance with guidelines for statin prescription—a multilevel methodological approach. BMC Health Serv Res 7:68

Raudenbush S, Bryk AS (1986) A hierarchical model for studying school effects. Sociol Educ 59:1–17

Renders CM, Valk GD, Franse LV, Schellevis FG, van Eijk JT, Van der Wal G (2001) Long-term effectiveness of a quality improvement program for patients with type 2 diabetes in general practice. Diabetes Care 24:1365–1370

Snijders TAB (2001) Sampling. In: Leyland AH, Goldstein H (eds) Multilevel modelling of health statistics. Wiley, Chichester

Stinchcombe AL (2005) The logic of social research. University of Chicago Press, Chicago

Sundquist K, Eriksson U, Kawakami N, Skog L, Ohlsson H, Arvidsson D (2011) Neighborhood walkability, physical activity, and walking behavior: the Swedish Neighborhood and Physical Activity (SNAP) study. Soc Sci Med 72:1266–1273

Twisk JWR (2006) Applied multilevel analysis. Cambridge University Press, Cambridge

Zegers M, De Bruijne MC, Spreeuwenberg P, Wagner C, Van der Wal G, Groenewegen PP (2011) Variation in rates of adverse events between hospitals and hospital departments. Int J Qual Health Care 23:126–133