

Chapter 6

Apportioning Variation in Multilevel Models



Abstract The starting point of multilevel analysis is to separate the variance in an outcome into the parts that are associated with the levels we distinguish. Several issues concerning variances at all levels are discussed in this chapter. Partitioning the variance between levels is straight forward in two-level linear models, but more complicated when we consider more than two levels or when our outcome is dichotomous. We discuss ways of clarifying and interpreting the importance of the higher level variance in logistic multilevel regression analysis. This can be done by transforming the difference between the 2.5 and 97.5 centile into an odds ratio or by using the median odds ratio, both of which can be interpreted in the same way as the odds ratio of a specific fixed effect. The higher level variance estimated from multilevel logistic regression models tends to be low leading to the question as to whether this small variance is still important. The clustering of observations within higher level units also informs power calculations. More variance at a higher level means that we need more observations to achieve the same power as when there is little variation at the higher level. The specification of levels is important, both from a theoretical and a statistical point of view. Omitting a relevant level has consequences for the estimation of the amount of variation associated with the remaining levels.

Keywords Multilevel analysis · Variance partitioning · Multilevel logistic regression · Median odds ratio · Power calculation · Design effect · Omitted level bias

One feature of multilevel models that is absent in single-level models is the ability to partition any unexplained variance between levels and hence quantify the importance of different levels. As we explained in Chap. 3, we can develop hypotheses solely concerned with variation in the phenomenon we are studying. In this chapter we give further consideration to the important topic of variance, and we consider interpretation of the variance and expound upon the implications of the variance both for model interpretation and for study design.

Variance Partitioning for Continuous Responses

In Chap. 5 we saw that the intraclass correlation coefficient ρ_1 was a simple summary of the proportion of the total variance in a two-level random intercept model that was attributable to the higher level.

$$\rho_1 = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2} \quad (6.1)$$

There are many situations in which the proportion of variance at a higher level cannot be summarised in such a simple fashion. These include circumstances when we have more than two levels (meaning that σ_{e0}^2 and σ_{u0}^2 are not the only variances), in the presence of heteroscedasticity (non-constant level 1 errors, in which case σ_{e0}^2 is not the only variance at level 1), and when we are fitting a model with random slopes (σ_{u0}^2 is not the only variance at level 2).

In general the proportion of the total variance that is attributable to a particular level in the model, for a given set of compositional and contextual characteristics, is called the variance partition coefficient (VPC; Goldstein et al. 2002). In many cases the VPC must be calculated for specific values of the covariates included in a multilevel regression model. For example, in the case of a two-level random slope model with a continuous outcome written as

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{0ij} \quad (6.2)$$

where the random part is given by $u_{0j} + u_{1j} x_{1ij} + e_{0ij}$ which depends on x_{1ij} . This means that the total variance, and therefore also the proportion of the variance that is at level 2, varies according to the value of the level 1 characteristic x_{1ij} .

Variance Partitioning for Multilevel Logistic Regression

In a multilevel logistic regression model, the VPC cannot be defined as in Eq. (6.1) even in the simplest variance component model. As detailed in Chap. 12, the observed outcome y_{ij} , a dichotomous response taking the value 1 if true and 0 otherwise, is modelled as a binomial process with denominator 1 and probability π_{ij} such that

$$y_{ij} \sim \text{Binomial}(1, \pi_{ij}) \quad (6.3)$$

In a random intercept model with a series of predictors x_{pij} , the transformed probability π_{ij} is modelled such that

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \cdots + u_{0j} \quad (6.4)$$

Now because of the assumption of a binomial distribution, the variance of the y_{ij} is given by $\pi_{ij}(1 - \pi_{ij})$. This is dependent on the predicted values π_{ij} and so, in turn, is dependent on all of the covariates x_{pij} . Moreover, the random effects u_{0j} , again assumed to be normally distributed with variance σ_{u0}^2 , are on the logit scale, and so it is not possible to make a direct comparison between the level 2 variance σ_{u0}^2 and the total variance $\pi_{ij}(1 - \pi_{ij})$.

Goldstein et al. (2002) discuss four approaches to the estimation of the VPC. The first approach, and the most commonly used, is the latent variable method used by Snijders and Bosker (2012, Chap. 17). This entails substituting the constant quantity $\pi^2/3 \approx 3.29$ for the lowest level variance, meaning that for a two-level multilevel logistic regression model with a random intercept,

$$\rho_1 = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \pi^2/3} \quad (6.5)$$

The second is a simulation method that is generalisable and has the advantage of not depending upon approximations. The third uses a Taylor series expansion (a power series approximation of a mathematical function) to provide an algebraic approximation for the VPC. The last method uses a binary linear model; this is a very approximate approach that involves treating the dichotomous responses y_{ij} as though they are normally distributed and fitting a model accordingly and, as such, tends to work better when the probability of the outcome is close to 0.5 rather than close to 0 or 1.

Variance Partitioning for Models with Three or More Levels

In the presence of more than two levels, the VPC details the proportion of unexplained variance that is attributable to the different levels in the model. Merlo et al. (2012) modelled the probability of death using a multilevel logistic regression model with four levels; individuals were nested within households, which were in turn clustered within census tracts and municipalities. The authors estimated variances associated with the three highest levels, denoted by σ_H^2 , σ_C^2 and σ_M^2 , respectively, and used these to calculate the VPCs under the latent variable method (Snijders and Bosker 2012) as

$$\begin{aligned} \text{VPC}_M &= \sigma_M^2 / (\sigma_M^2 + \sigma_C^2 + \sigma_H^2 + \pi^2/3) \\ \text{VPC}_C &= (\sigma_M^2 + \sigma_C^2) / (\sigma_M^2 + \sigma_C^2 + \sigma_H^2 + \pi^2/3) \\ \text{VPC}_H &= (\sigma_M^2 + \sigma_C^2 + \sigma_H^2) / (\sigma_M^2 + \sigma_C^2 + \sigma_H^2 + \pi^2/3) \end{aligned} \quad (6.6)$$

Note that these variance partition coefficients are cumulative, indicating the proportion of unexplained variance at the level in question and at higher levels. This means that they can also be interpreted as the correlation between individuals from the same higher level unit; individuals living in the same household must live in the same census tract and people from the same census tract must live in the same municipality since these are strictly clustered. It is straight forward to calculate the proportion of the total variance associated with a particular level by subtraction. For example, in the null model, estimates of VPC_H and VPC_C were 0.186 and 0.023, respectively, indicating a correlation in mortality between individuals within the same household of 0.186 and suggesting that 16.3% of the total variance in mortality was attributable to differences between households within census tracts.

Interpretation of Variances

In a multilevel model with a random intercept, the interpretation of the variance in terms of the VPC—however estimated—is fairly straightforward. For example, Gonzalez et al. (2012) investigated the clustering of young adults' body mass index (BMI) within families; for a two-level null model, they reported a variance between families (σ_{u0}^2) of 8.92 and a variance between young adults within families (σ_{e0}^2) of 13.92. The variance partition coefficient (which in this simple case is the same as an intraclass correlation coefficient) is therefore

$$VPC = \sigma_{u0}^2 / (\sigma_{u0}^2 + \sigma_{e0}^2) = 8.92 / (8.92 + 13.92) = 0.391$$

This means that they found 39.1% of the variation in BMI in young adulthood to be attributable to the family level, with the remaining 60.9% being due to differences between young adults within families. The total variance in the sample is 22.84 and so the standard deviation σ is 4.779; with a reported mean BMI of 25.38, we would expect 95% of the young adults to have a BMI of between $(\mu - 1.96\sigma, \mu + 1.96\sigma) = (16.01, 34.75)$. We can also say something about the variation between families; we would expect 95% of families to have a mean young adult BMI of between $(\mu - 1.96\sqrt{\sigma_{u0}^2}, \mu + 1.96\sqrt{\sigma_{u0}^2})$ or (16.99, 28.69).

In multilevel logistic regression models, we have less information available—just the higher level variance σ_{u0}^2 in a two-level random intercept model—and our interpretation of the variance is different. We are, however, still able to interpret the random part of a multilevel logistic regression model, and given that it is slightly more complex, this is arguably more important than for the multilevel linear regression model. For example, Esser et al. (2014) examined in-hospital mortality among very low birthweight neonates in Bavaria. Following adjustment for individual casemix (including gestational age, sex and the clinical risk index for babies

[CRIB] score), the authors found a variance between hospitals (σ_{u0}^2) of 0.324. Assuming the latent variable method discussed above, the variance partition coefficient calculated according to Eq. (6.5) is given by

$$\text{VPC} = 0.324 / (0.324 + 3.29) = 0.090$$

In other words, 9.0% of the total variation in mortality is attributable to differences between hospitals after adjustment for casemix (with the remaining 91.0% relating to differences between patients within hospitals that have not been accounted for by variables included in the model). The high-level variance σ_{u0}^2 is again informative, but this time it is on a log odds scale. We would expect 95% of hospitals to have a log odds ratio of mortality—relative to the typical hospital—of $(-1.96\sqrt{\sigma_{u0}^2}, +1.96\sqrt{\sigma_{u0}^2})$. Converting this to an odds ratio scale (by exponentiating), we would expect 95% of hospitals to have an odds ratio (OR) of mortality associated with being in that hospital, compared to the typical hospital, to be in the interval $(\exp\{-1.96\sqrt{\sigma_{u0}^2}\}, \exp\{1.96\sqrt{\sigma_{u0}^2}\})$ or (0.33, 3.05).

Rather than considering the 95% coverage intervals, we can make comparisons between the upper and lower limits of the distribution. Returning to the example of Gonzalez et al. (2012), we would expect the mean BMI of a family at the 97.5 centile to exceed that of a family at the 2.5 centile by $2 \times 1.96\sqrt{\sigma_{u0}^2} = 11.71$ —the difference, save for rounding error, between the upper and lower 95% limits of 28.69 and 16.99 calculated above. So 95% of families should be covered by 11.71 points on the BMI scale. It is possible to do something similar for a logistic regression model; we would expect the odds of mortality for a hospital at the 97.5 centile to be $\exp\{2 \times 1.96\sqrt{\sigma_{u0}^2}\}$ or 9.31 times the odds of mortality associated with a hospital at the 2.5 centile. Again, apart from rounding error, this is approximately the ratio of the two limits of the coverage interval, 3.05 and 0.33.

The variance estimate from a multilevel logistic regression model can therefore be used as a means of informing us about the variation between higher level units in the dataset. The comparison of the 97.5 and 2.5 centiles is arbitrary; a commonly used alternative that is not dependent on such an arbitrary range and which was introduced by Larsen and Merlo (2005) is the median odds ratio (MOR). The MOR is the median of odds ratios comparing two people with identical covariates chosen randomly from different higher level units (ordered so that the odds ratio is always at least one). It is calculated as

$$\text{MOR} = \exp\left\{\sqrt{2 \times \sigma_{u0}^2} \Phi^{-1}(0.75)\right\} \quad (6.7)$$

$\Phi^{-1}(0.75)$ is the 75th centile of the standard normal density or 0.6745 giving

$$\text{MOR} = \exp \left\{ 0.954 \times \sqrt{\sigma_{u0}^2} \right\} \quad (6.8)$$

In the example of in-hospital mortality among very low birthweight neonates given by Esser et al. (2014), the variance of 0.342 gives an MOR of 1.72. This calculation has converted the variance to a measure of dispersion on the odds ratio scale, telling us something about the average difference between two random hospitals. Since it is now on the odds ratio scale, this can be compared to any other odds ratio, for example, for any of the fixed effects such as sex.

The MOR is used as a means of transforming the variance onto a meaningful and interpretable scale in multilevel logistic regression; there are equivalent measures for other forms of multilevel analysis. Chan et al. (2011) found a median rate ratio (MRR) of 1.31 between practices for treatment using warfarin among patients with nonvalvular atrial fibrillation using a multilevel modified Poisson regression model. Chaix et al. (2007) reported a median hazard ratio (MHR) of 1.25 between small areas in Sweden when analysing ischaemic heart disease mortality. The calculation of the MRR and the MHR follows the same principles as for the MOR. More details about the MRR can be found in Austin et al. (2018) and details of the MHR in Austin et al. (2017).

The MOR and related measures make use of the distributions of the residuals and are easy to calculate since they depend only on the higher level variance σ_{u0}^2 . An alternative measure, the absolute relative deviation (ARD), quantifies the average difference between the effect of each high-level unit and the effect of an average high-level unit (see Martikainen et al. 2003; Tarkiainen et al. 2010). The ARD uses the model residuals u_{0j} and so is more complicated to calculate but may be particularly useful when there are fewer higher level units (and the distribution of these higher level units may not strictly follow a standard statistical distribution).

Zero Variance

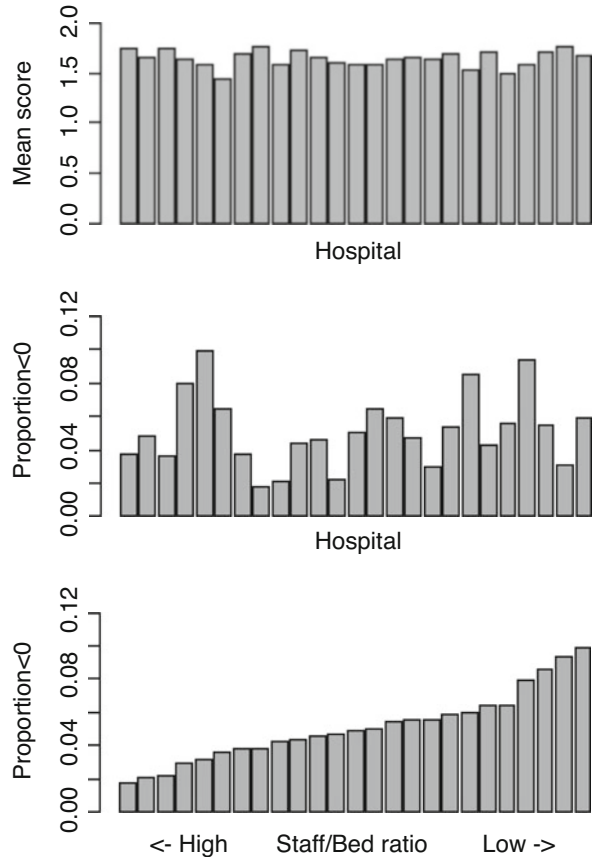
Unexplained variance between high-level units may constitute a small proportion of the total variance in the outcome. Unfortunately there is no consensus as to exactly what constitutes a ‘small’ proportion. Usually the higher level variance is small compared to the lower level variance. The common exception is for repeated measures in which there will typically be less variability between measurement occasions than between the higher level units. For example, in a study of health functioning in a cohort of British civil servants, Stafford et al. (2008) found 57% of the variation in physical functioning and 49% of the variation in mental functioning at baseline to be associated with the level of the individual. Chapter 11 describes the modelling of repeated measures on areas rather than individuals; in that example 82% of the variation in mortality rates is seen to be at the district level.

In some situations, a higher level variance will be estimated to be zero. The suggestion that all of the unexplained variation is at the individual (lowest) level does not mean that the mean outcome is identical for all contexts; rather, this means that there is no more variation between higher level units than we would have expected by chance. But that does not mean that there is no variation, and, at first sight, the differences between high-level units may appear substantial.

To illustrate this concept, we simulate a random assignment of individuals to 25 hospitals, with each hospital comprising between 90 and 120 patients. Each patient has a ‘vitality score’; these scores are generated as random draws from a normal distribution with mean 1.64 and variance 1. Figure 6.1a shows the mean scores for the 25 hospitals under one such simulation. There is little variation between the hospital means—the minimum and maximum are 1.45 and 1.77 with the variance of the hospital mean scores (0.007) being very small compared to the individual variance of 1 that was used to generate these data.

If the vitality score is such that a patient with a score of 0 or more denotes life and a score below 0 denotes death, then we can use the individual scores to categorise

Fig. 6.1 Simulated data of individuals aggregated to hospitals showing (a) the variation between hospitals in the mean vitality score; (b) the variation between hospitals in the mortality rate (for whom the vitality score <0) and (c) the association between the hospital mortality rate and a contextual variable



patients. A score of 0 corresponds to -1.64 standard deviations so about 5% of all patients will be classified as dead. Figure 6.1b shows the results of aggregating the individual patient deaths to the hospital level and expressing these as a proportion. The proportion of deaths in each hospital now ranges between 0.018 and 0.099, but this fivefold difference in mortality rates between hospitals has occurred by chance. We would quite reasonably estimate the variance between hospitals to be zero since there is no more than we would expect by chance.

When there is no variance between higher level units in a two-level model, the intraclass correlation coefficient is 0 and the model effectively collapses to a single-level model. However, in such circumstances Merlo et al. (2009) point out that this should not exclude the possibility of investigating (and indeed discovering) contextual effects. Figure 6.1c shows how the ranking of hospitals in terms of their mortality rates may be correlated with key staffing indicators such as the staff/bed ratio. Despite there being no unexplained variance being associated with the hospitals, we can find a significant relationship with a contextual variable.

Merlo et al. (2012) argue that the general contextual effects (the overall extent to which context influences individual health outcomes, assessed using the variance and VPC) should have greater prominence in research and that such measures are more informative than tests of the significance of small area variation common in spatial epidemiological analysis. The authors further suggested that the small variances typically found at the area level should lead to less importance being placed on administrative areas as a determinant of individual health than is currently the case (see also our discussion of the relevance of contexts in Chap. 2).

Multilevel Power Calculations

Power calculations are an important aspect of study designs involving primary data collection and are often regarded as essential by funders even for the analysis of existing data. When a study includes different levels, it is necessary to take these into account when conducting the power calculation; failure to do so will lead to an overestimation of the power available for the analysis since the lack of independence between observations nested within the same higher level unit reduces the effective sample size.

The focus of the power calculation depends on its purpose. Common uses are to indicate the power that is available to detect a specified effect with a given sample size, the sample size needed to detect a specified effect at a given level of power or an estimate of the effect size that could be detected with a given sample size at a specified level of power. The three quantities power, sample size and effect size are related, and so the unknown quantity can be changed by simple algebraic manipulation. (We have assumed that the significance level used is the common $\alpha = 0.05$.) As is the case for single-level power calculations, two of the three quantities are assumed to be known in order to estimate the third. However, specifying the sample size in a multilevel design is more complicated; in addition to the number of

individuals (level 1 units), we need to know the number of level 2 units and the extent of the clustering of the outcome within the level 2 units as expressed by the intraclass correlation coefficient ρ_1 .

The calculation of the required sample size n for a single-level problem with power β to detect an effect size of magnitude x/σ at a significance level α is as follows:

$$n = \left[\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{x/\sigma} \right]^2 \quad (6.9)$$

Z_r is the value from the standard normal distribution with the proportion r below it, and so $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96$. The effect size here is standardised and expressed in terms of the number of standard deviations and assumes that the outcome is normally distributed; equivalent formulae are available when the dependent variable is dichotomous.

The multilevel data structure is taken into account by inflating the variance in Eq. (6.9) by a design effect D

$$D = 1 + (\bar{n}_j - 1)\rho_1 \quad (6.10)$$

\bar{n}_j is the average number of individuals (level 1 units) in a cluster. The design effect therefore depends on both the magnitude of the intraclass correlation coefficient and the average cluster size. If $\rho_1 = 0$, there is no correlation between individuals within the same high-level unit, $D = 1$, and the power is the same as for a simple random sample. If $\rho_1 = 1$ there is no variation within high-level units, $D = \bar{n}_j$, and there is no gain through sampling more than one individual per cluster. Power can only be increased in this instance by sampling more level 2 units. If $\bar{n}_j = 1$, then only one individual is being sampled per cluster, $D = 1$, and the power is the same as for a simple random sample. In general, D will be greater than 1 and the clustering of outcomes within contexts reduces the power of a multilevel model relative to a simple random sample.

The dependence of the power calculation on the design effect means that we need to have an idea of the likely magnitude of the design effect. Design effects can often be calculated based on the reporting of intraclass correlation coefficients in the literature. For example, if we were interested in compliance with a colorectal cancer screening programme, we might base our power calculation on the study by Pernet et al. (2011). They found a variance between geographical areas in France (Ilôts Regroupés pour l'Information Statistique, IRIS) of 0.040 in an empty model. Given that this estimate was derived from a multilevel logistic regression model, the estimated intraclass correlation coefficient calculated using Eq. (6.5) is 0.012. This means that an estimated 1.2% of the variation in uptake of screening is associated with the area. This study was based on the analysis of 8691 individuals in 829 IRISs; if we were to take a similar sample, then the average cluster size would be $\bar{n}_j = 10.48$. Based on Eq. (6.10), the design effect is given by

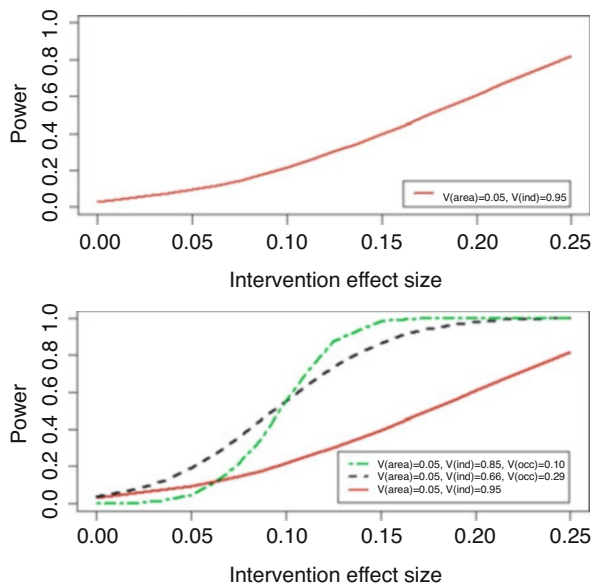
$$D = 1 + (10.48 - 1) \times 0.012 = 1.11$$

Even with a trivial intraclass correlation coefficient, and a modest average cluster size, the clustering of individuals within areas means that we would need to increase our sample size by 11% to get the same power as a simple random sample of uncorrelated individuals. Note that this increase in sample size needs to be reflected in an increase in the number of areas, since an increase in the number of individuals per area would in turn increase the magnitude of the design effect.

It is possible that a literature search will turn up a relevant research article from which the intraclass correlation coefficient can be found for a multilevel power calculation. There are also resources reporting intraclass correlation coefficients for different study types, such as those for various health outcomes in UK settings (Ukoumunne et al. 1999), cardiovascular disease in primary care practices in Canada (Singh et al. 2015) and BMI, physical activity and diet across countries (Masood and Reidpath 2016). The need for information to perform power calculations is a further argument for the need to report the intraclass correlation or variances in research articles (see Chap. 10 for further discussion of this).

From the above, it would appear that a large intraclass correlation coefficient is the enemy of efficient and economical study design, with even small intraclass correlation coefficients leading to substantial increases in the sample size required (and hence in many cases, the cost involved) to replicate the power of a simple random sample. However, this is design dependent since a repeated measure design—with a large associated intraclass correlation coefficient—can increase the power of an analysis. We can illustrate this by considering two simulated study designs for the evaluation of an area intervention. Figure 6.2a shows how the power available to detect a specific effect size increases with the effect size in a *repeated*

Fig. 6.2 Simulation showing the power to detect an effect of given size for an area-based intervention based on (a) a repeated cross-sectional design (40 individuals in 50 areas) and (b) a repeated measures design (two observations on 20 individuals in 50 areas)



cross-sectional design. This simulated study has 20 individuals measured before the intervention and 20 after the intervention in each of 50 areas, assuming an intraclass correlation coefficient of 0.05. With this design, the effect size has to be close to 0.25 before the power reaches 0.8. In Fig. 6.2b, the study design is changed to a *repeated measures* design, such that each of 20 individuals is measured before and after the intervention (two measurements per person). This design retains the same total number of measurements as in the repeated cross-sectional design (2000), and the total variance is unchanged, but there is now some variation within as well as between individuals. This study is now more highly powered to detect effects of modest sizes, with power of 0.8 to detect an effect size of 0.11–0.13 based on the same proportion of the variance at the area level as in Fig. 6.1a but with 69–89% of the remaining variance being attributable to differences between individuals. With this study design, the fact that a relatively small proportion of the total variance (10–29%) is associated with the measurement occasion means that any change between the pre- and post-intervention measures is more likely to denote an effect of the intervention.

Power calculations are commonly used to determine whether it will be possible to detect an effect of a certain size; as such they involve the comparison of the magnitude of a parameter estimate to its precision (as measured by its standard error). But the accuracy of different parameter estimates, and their standard errors, may also 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) level 2 units. With fewer level 2 units, the only estimate that was affected was the standard error of the high-level variance.

Software for Multilevel Power Calculations

In the simplest designs, it may be possible to inflate the sample size required over that needed for a simple random sample using the design effect, as we did for the example on compliance with colorectal cancer screening above. However, this may not be straightforward for more complicated designs such as when there is considerable lack of balance between cluster sizes or when the effect size to be estimated is not at the lowest level (such as the simulated area-based intervention above). For such circumstances, specialist software is available, for example, MLPowSim (Browne et al. 2009), Optimal Design (Spybrook et al. 2011) and PINT (Snijders and Bosker 1993). The topic along with the software has also been covered in some detail by Moerbeek and Teerenstra (2015).

There may be other constraints on the sample size calculation such as cost. In particular, cost may be an important consideration when there is a cost associated with each higher level unit that is sampled over and above the costs of the individuals sampled. This is the case if, for example, we had to organise data collection in more hospitals, needing permissions, time of hospital personnel, field workers, etc. Snijders (2001) gives an example of incorporating cost considerations into a multilevel study design.

Population Average and Cluster-Specific Estimates

The parameter estimates obtained from a multilevel model are sometimes called *cluster-specific* (or *random effect*) estimates. These estimates are conditional on the random part of the model and therefore indicate the effect of the variable in question on two individuals from the same higher level unit. In contrast, *population average* (also called *marginal*) estimates indicate the effect of a covariate on the average person (Diez-Roux 2002). The two estimates are identical for normally distributed responses but will tend to differ for non-linear responses, such as for a logistic regression model, with the differences becoming larger as the variance increases. Population average estimates are usually given as the output from generalised estimating equations (GEEs; see Zeger et al. 1988) whilst cluster-specific estimates are the default output from most multilevel modelling packages. The population average estimate β^* is approximately related to the cluster-specific estimate β as follows (Larsen and Merlo 2005):

$$\beta^* \approx \beta / \sqrt{1 + 0.346\sigma_{u0}^2} \quad (6.11)$$

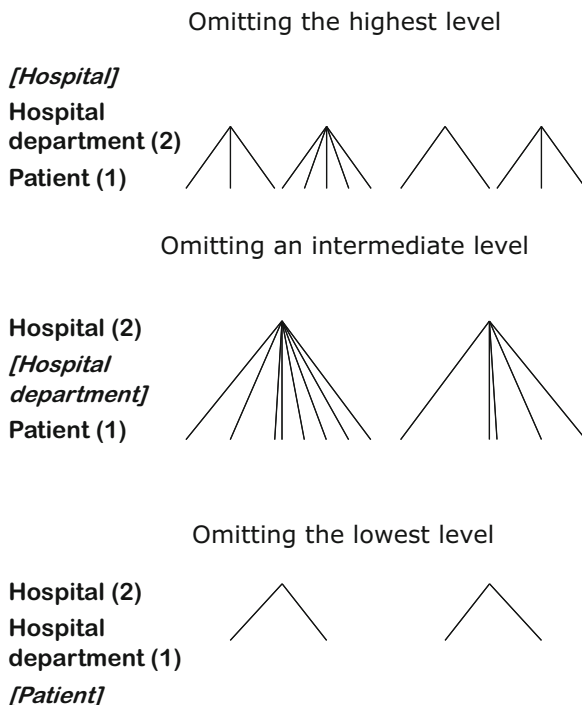
Note that β and β^* here are parameter estimates on their original scale, i.e. log odds ratios for a logistic regression model. As can be seen from Eq. (6.11), the smaller the variance σ_{u0}^2 the smaller the difference between the two estimates. For example, with an estimate of $\beta = 1.40$ (giving an odds ratio OR = 1.49), a variance of $\sigma_{u0}^2 = 0.05$ leads to a population average estimate of $\beta^* = 0.397$ (OR = 1.49) whilst a variance of $\sigma_{u0}^2 = 0.50$ gives $\beta^* = 0.369$ (OR = 1.45). So if required (e.g. if requested by a journal), population average effects can be obtained from the cluster-specific effects. The distinction between the multilevel and GEE approaches is explored in more detail elsewhere (Burton et al. 1998; Hu et al. 1998; Hubbard et al. 2010).

Omitting a Level

Suppose we have a study which has data on two levels. A theoretical analysis of our research problem might lead us to hypothesise the importance of other levels too. What is the consequence of omitting a theoretically important level for the interpretation of the partitioning of variance? A statistical and empirical analysis (using UK Census data) was made by Tranmer and Steel (2001). We distinguish between three situations shown in Fig. 6.3.

To make it more concrete, think of an example in which we are studying episodes of care of patients admitted to hospital departments. The data we have in Fig. 6.3a refer to patients within hospital departments. It may also be important to have information on the hospitals. In Fig. 6.3b, we have data on patients and hospitals, but not on the hospital department. Finally, in Fig. 6.3c, we lack any information at the level of the patient.

Fig. 6.3 Illustration of omitting different theoretically important levels



What happens to the variance in these situations? The first situation is quite straightforward (although not satisfactory); the variation at the highest level in Fig. 6.3a is combined with the variation at the next level down and is indistinguishable from it. In the example of hospitals, the variance estimated at the level of the department includes variance between hospitals as well as variance between departments within hospitals, but we do not know the proportion of the variance that is associated with each of these two levels. The patient-level variance will, however, be estimated accurately. In Fig. 6.3b, the department level is omitted, and the associated variance is distributed between the patient and hospital levels. Sacker et al. (2006) give an example of such a situation. They studied self-rated health of individuals taken from the British Household Survey at different times. They compared a model with two levels, individuals nested within areas (electoral wards) and a model with three levels where the level of the household is included between individuals and areas. As Fig. 6.4 shows, part of the individual-level variance estimated from the two-level model is actually related to the households people live in and (a smaller) part of the area-level variance estimated from the two-level model turns out to be associated with variation between households within the areas.

Tranmer and Steel (2001) show that for a linear model the proportion of the intermediate-level variance that will be distributed to the highest level is approximately \bar{n}_{jk}/\bar{n}_k , where \bar{n}_{jk} and \bar{n}_k are the average cluster size (in terms of level 1 units, e.g. individuals) at the intermediate and highest levels, respectively, with the remainder being distributed to the lowest level. However, if the magnitude of the

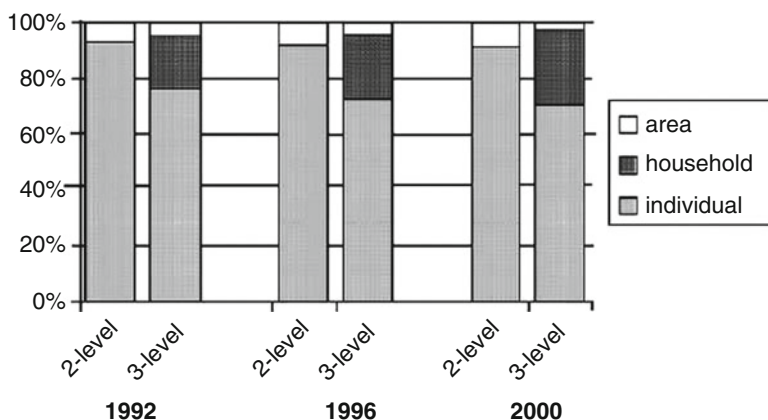


Fig. 6.4 Proportion of variance at each level of a two-level (individuals within areas) and three-level (individuals within households within areas) baseline model of poor general health in the British Household Panel Study. (Reproduced with permission from Elsevier, Health & Place)

variance at the omitted level is unknown, it is impossible to assess the impact of its omission.

When the lowest level is omitted as in Fig. 6.3c, the model is rather different since the analysis is aggregated to the intermediate level (such as hospital department). The variance at the highest level (hospital) is estimated correctly, but the estimated variance at the intermediate level (department) includes a component from the lowest (individual) level. Although the proportion of the lowest level variance that is incorrectly attributed to the intermediate level is likely to be small—Tranmer and Steel (2001) estimate this proportion to be just $1/\bar{n}_{jk}$ —this will commonly be a small proportion of a large variance since σ_{u0}^2 is commonly much smaller than σ_{e0}^2 . For example, let us assume that in the correctly specified three-level model, 5% of the variance is at the level of the hospital, 5% at the level of the hospital department and the remaining 90% of the variance refers to differences between patients within hospital departments. The fact that σ_{e0}^2 is 18 times σ_{u0}^2 means that, even if there are as many as 100 patients in each hospital department ($\bar{n}_{jk} = 100$), omitting the patient level would result in an 18% inflation of the estimated variance between hospital departments.

Conclusion

The variances at different levels form an important and informative part of the multilevel model and even small variances at higher levels can have a substantial impact on the required sample sizes. Despite their importance for model interpretation, assessment of the importance of contexts and for the conduct of future power calculations, Riva et al. (2007) found in a review that many studies did not report variance components. This is clearly an oversight by authors and journals, and we

would hope that this situation will improve over time. In Chap. 10 (Reading and writing), we further emphasise the importance of reporting variances from multilevel studies. We have also seen that when a level is omitted from an analysis, the impact on the variances estimated in the (incorrectly specified) multilevel model is unpredictable.

References

- Austin PC, Wagner P, Merlo J (2017) The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis. *Stat Med* 36:928–938
- Austin PC, Stryhn H, Leckie G, Merlo J (2018) Measures of clustering and heterogeneity in multilevel Poisson regression analysis of rates/count data. *Stat Med* 37:572–589
- Browne WJ, Lahi MG, Parker RMA (2009) A guide to sample size calculations for random effect models via simulation and the MLPowSim Software Package. School of Clinical Veterinary Sciences, University of Bristol, Bristol
- Burton P, Gurrin L, Sly P (1998) Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 17:1261–1291
- Chaix B, Rosvall M, Merlo J (2007) Assessment of the magnitude of geographical variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. *J Epidemiol Community Health* 61:349–355
- Chan PS, Maddox TM, Tang F, Spinler S, Spertus JA (2011) Practice-level variation in warfarin use among outpatients with atrial fibrillation (from the NCDR PINNACLE program). *Am J Cardiol* 108:1136–1140
- Diez-Roux AV (2002) A glossary for multilevel analysis. *J Epidemiol Community Health* 56:588–594
- Esser M, Lack N, Riedel C, Mansmann U, von Kries R (2014) Relevance of hospital characteristics as performance indicators for treatment of very-low-birth-weight neonates. *Eur J Public Health* 24:739–744
- Goldstein H, Browne W, Rasbash J (2002) Partitioning variation in multilevel models. *Understanding Stat* 1:223–231
- Gonzalez A, Boyle MH, Georgiades K, Duncan L, Atkinson LR, MacMillan HL (2012) Childhood and family influences on body mass index in early adulthood: findings from the Ontario Child Health Study. *BMC Public Health* 12:755
- Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA (1998) Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *Am J Epidemiol* 147:694–703
- Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, Bruckner T, Satariano WA (2010) To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 21:467–474
- Larsen K, Merlo J (2005) Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol* 161:81–88
- Maas CJM, Hox JJ (2005) Sufficient sample sizes for multilevel modeling. *Methodology* 1:86–92
- Martikainen P, Kauppinen TM, Valkonen T (2003) Effects of the characteristics of neighbourhoods and the characteristics of people on cause specific mortality: a register based follow up study of 252 000 men. *J Epidemiol Community Health* 57:210–217
- Masood M, Reidpath DD (2016) Intraclass correlation and design effect in BMI, physical activity and diet: a cross-sectional study of 56 countries. *BMJ Open* 6:e008173
- 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

- Merlo J, Viciana-Fernández FJ, Ramiro-Fariñas D, the Research Group of the Longitudinal Database of the Andalusian Population (LDAP) (2012) Bringing the individual back to small-area variation studies: a multilevel analysis of all-cause mortality in Andalusia, Spain. *Soc Sci Med* 75:1477–1487
- Moerbeek M, Teerenstra S (2015) Power analysis of trials with multilevel data. Chapman & Hall, Boca Raton
- Pornet C, Dejardin O, Morlais F, Bouvier V, Launoy G (2011) Socioeconomic determinants for compliance to colorectal cancer screening: a multilevel analysis. *J Epidemiol Community Health* 64:318–324
- Riva M, Gauvin L, Barnett TA (2007) Towards the next generation of research into small area effects on health: a synthesis of multilevel investigations published since July 1988. *J Epidemiol Community Health* 61:853–861
- 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
- Singh J, Liddy C, Hogg W, Taljaard M (2015) Intraclass correlation coefficients for sample size calculations related to cardiovascular disease prevention and management in primary care practices. *BMC Res Notes* 8:89
- Snijders TAB (2001) Sampling. In: Leyland AH, Goldstein H (eds) *Multilevel modelling of health statistics*. Wiley, Chichester
- Snijders TAB, Bosker RJ (1993) Standard errors and sample sizes in two-level research. *J Educ Stat* 18:237–260
- Snijders TAB, Bosker RJ (2012) *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage, Los Angeles
- Spybrook J, Bloom H, Congdon R, Hill C, Martinez A, Raudenbush S (2011) Optimal design plus empirical evidence: documentation for the “Optimal Design” software version 3.0. Available from <http://www.wtgrantfoundation.org/resources/optimal-design>
- Stafford M, Gimeno D, Marmot MG (2008) Neighbourhood characteristics and trajectories of health functioning: a multilevel prospective analysis. *Eur J Public Health* 18:604–610
- Tarkiainen L, Martikainen P, Laaksonen M, Leyland AH (2010) Comparing the effects of neighbourhood characteristics on all-cause mortality using two hierarchical areal units in the capital region of Helsinki. *Health Place* 16:409–412
- Tranmer M, Steel DG (2001) Ignoring a level in a multilevel model: evidence from UK census data. *Environ Plan A* 33:941–948
- Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ (1999) Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 3(5):iii-92
- Zeger S, Liang K-Y, Albert P (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049–1060

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

