Bachelor's project

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Louise Bauer-Nilsen

Analysing Nested Data with Multilevel Models

Bachelor's project in Mathematical Sciences

Supervisor: Geir-Arne Fuglstad

May 2020

NTNU
Norwegian University of Science and Technology

Louise Bauer-Nilsen

Analysing Nested Data with Multilevel Models

Bachelor's project in Mathematical Sciences
Supervisor: Geir-Arne Fuglstad
May 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

For å analysere nestede datasett bruker vi metoden for flernivåmodellering. Med nestede datasett mener vi at vi har en naturlig gruppering i datasettet. Ved å bruke flernivåmodellering til å analysere dataen kan vi få innsikt i variasjon mellom og innad i grupper og forskjellen i variabilitet for de forskjellige flernivåmodellene. Vi bruker R-pakken lme4 og bruker funksjonen lmer() for å tilpasse de forskjellige flernivåmodellene. For de forskjellige modellene lar vi enten skjæringspunktet eller stigningstall variere. Dette gir oss tre forskjellige modeller, modellen med varierende skjæringspunkt, modellen med varierende stigningstall, og til sist, modellen hvor vi lar både stigningstall og skjæringspunkt variere. På modellene kan vi også legge til prediktorer på de forskjellige nivåene. Vi bruker et eksempel på radondata der vi vil måle radonnivåene i amerikanske hjem, dataen er strukturert slik at vi har husholdninger inni de forskjellige fylkene. Vi tilpasser dataene våre og kan se, etter modell sjekking, at en mer komplisert modell vil forbedre tilpasningen, og vi kan også observere variansen i de forskjellige nivåene.

Abstract

To analyse nested data we use the method of multilevel modeling. By nested data we mean naturally structured groups within our data. With the information we receive by fitting multilevel models, we can illustrate how the group variances and the model variability change for the different multilevel models and how to analyse the output from R code. We use the R package lme4 and use the function lmer() to fit the different multilevel models. For the different models we allow either the intercept or slope of the model to vary. This gives us three different models, the random intercept model, the random slope model, or by allowing both to vary: random intercept, random slope model. For the different models we are able to add predictors at the different levels. We use an example of radon data where we want to measure the radon levels in US homes, the data is structured so that we have households nested within counties. We fit our data and observe with model checking that the model fit will improve with more complex models and calculate the explained variance at each level.

Contents

1 Introduction

We have a dataset consisting of nested data. This means that we have data that is naturally nested or grouped. Naturally nested data often occur in the social sciences, as we deal with people, geographic locations or environments with people. How come data depending on these factors become nested? Because this kind of data is more complicated, in an environment within a school, different factors can yield different outcomes. Students within the same class with the same teacher and same teaching environment will likely perform more similarly than students from another class. Or students with similar socioeconomic backgrounds can be more similar than children from different socioeconomic backgrounds. The importance of the predictors or factors will depend on what our outcome will be. If we look at test scores from a test performed by students in different classes within different school districts, we would assume that students within same school districts performs similarly, due to the economic status of the school district or that students within a class perform similarly depending on the style of the teaching. This is an example of a nested model which also shows that we can have predictors at each level. Throughout the text we will describe the different levels as the individual-level or level-1 (for individuals within groups) and the groups as group-level or level-2.

Choosing a multilevel model allows us to introduce group-effects in the model that capture between-group variation. The multilevel model allows us to estimate group averages as fixed effects and between-group variation as a random effect. We assume that the random effects are independent and identically distributed according to a normal distribution. The multilevel model can then share information between the different groups, which can improve the prediction for groups with small sample sizes. The most basic multilevel model is the random intercept model, which has a different intercept in each group. The intercept consists of a random part drawn from the normal distribution and a joint intercept as a fixed effect. This gives each county a different intercept, but each county has the same slope.

A more complicated model is the random slope model. The random slope model allows each group to have a seperate slope. This is achieved by letting the covariate have a different effect within the different groups by including a random effect to the slope parameter.

Throughout the text I will use the example of home radon measurement and remediation from Gelman and Hill (2007). They estimate the distribution of radon levels in approximately 3000 counties in the United States. The data has a multilevel structure: houses within counties. We have two important predictors:
1. An individual-level predictor: in which floor the measurement was taken (first floor or basement). Necessary as the ground is the source of radon, the closer we measure to the soil, the higher the level of radon.
2. Group-level predictor, the measurement of soil uranium at county-level. It is assumed that the radon level in the ground is similar for houses within the same county.
A multilevel model lets us fit a regression model to all measurements that takes into account the variation between the 3000 counties, houses and measurements (first-floor or basement). I have used the datasets which are used in Gelman and Hill (2007) for the radon datasets, which I have implemented into R. In Gelman and Hill (2007), they have used Bugs but this is outdated, today we would have used STAN instead (Carpenter et al., 2017).

Multilevel modeling, can be seen as a form of partial pooling, i.e as a compromise between no-pooling, which fits a separate regression line for each group and complete pooling, which fits the same regression line to each group. Complete-pooling ignores between- group variation and gives the same estimate for each group. No-pooling overfit the data and overestimates the between-county variation. In Figure 1 we can observe the over-fitting from no-pooling, and how complete pooling ignores within-county variation. We have fitted data to log radon for the all 85 counties in Minnesota, and chosen eight to display the different levels from different counties in the US. The amount of pooling will depend on the group-level variance and the number of houses within each group.

In the first section, we introduce the most used multilevel model, the random intercept model. We then go on to the more complicated model, the random slope, random intercept model. At last we will discuss model checking for the different models and examples we have used throughout the text. Easy concepts can become more difficult in multilevel modelling, such as explained variance. Variance is now measured at different levels, and

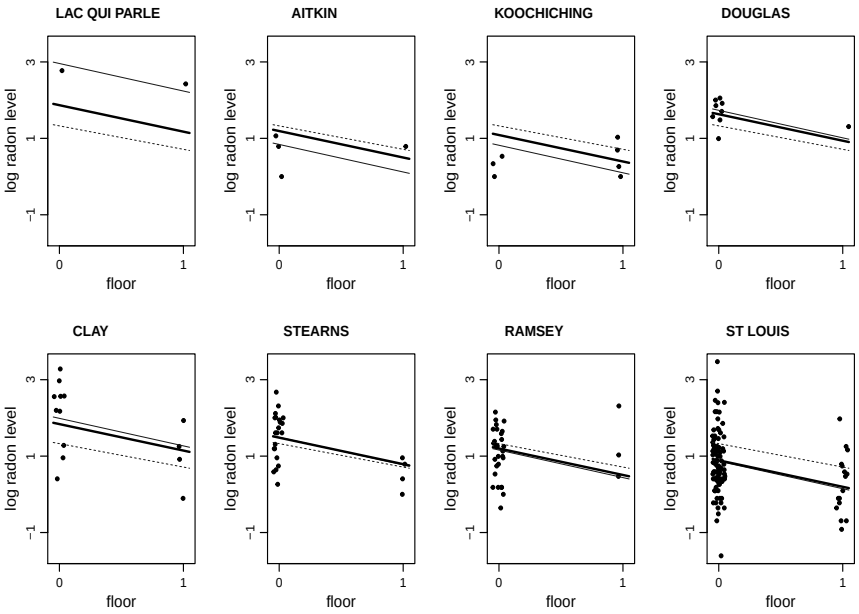


Figure 1: Regression fit to radon data from all the 85 counties in Minnesota. Multilevel (partial pooling) regression seen as the solid black line, no-pooling as the thin line and thin dashed line is the complete pooling regression.

common methods such as R-squared must be calculated at each level of the model, we will review this in the model checking section. The thesis ends with a discussion in section 5.

2 Random Intercept Model

The random intercept model is said to be the easiest version of the multilevel model, as well as being one of the most used multilevel models. A random intercept model is a model where we allow the intercept of the different groups to vary. Some of the groups have a higher response Y, and some have a lower response, creating different regression lines for different groups. For all multilevel models, we have combinations of random and fixed effects, and in the random intercept model we have the intercept as the random effect to make the intercept vary between the different groups. For this model the slope parameter will stay fixed. We repeat the information from the introduction, where we assume that we have predictors at each level in the multilevel model. j is the index for the groups, where $j = 1, ..., N$. i is the index for the individuals within the groups, $i = 1, ..., n_j$. Y_{ij} is the response variable and the variables within the model will depend on the indexes i and j, so that the notation x_{ij} describes the predictor for an individual i in group j. As the individuals i are nested within group j, its natural that the index i is always accompanied with the group index j. The random intercept model is given by

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 Z_j + \varepsilon_{ij} \\ \text{for } i = 1, ..., n_j. \\ \text{and for } j = 1, ..., N. \end{gathered} \tag{1}$$

where the β's are the regression parameters: β_{0j} is the intercept and β_1, β_2 are the fixed slope coefficients. x_{ij} is the predictor at the individual-level and the group-level predictor is given by z_j. ε_{ij} is namely the error at individual level and is normally distributed with a zero mean and variance σ^2_ε giving that all residuals are independent with variance constant across the groups.

We let the intercept vary between groups by adding a random part to the intercept. We split the group intercept into two parts, the mean overall intercept, γ_{00}, and the random variable U_{0j}, giving the unexplained group effects, also called the group-residuals. U_{0j} is a independent identically distributed random variable drawn from a normal distribution with mean zero and variance τ^2_0.

$$\beta_{0j} = \gamma_{00} + U_{0j} \tag{2}$$

This model contains unexplained variability at two nested levels. One of the main purposes of multilevel modeling is having a model for the response variable Y that takes

into account both an individual- and a group level variation. Thus taking the nested structure of the data into account. By looking at Example 1, we can see the random intercept model used in an example.

Example 1 Home Radon Measurements

For the home radon measurement we use Model (1), the random intercept model, where the response variable Y_{ij} is the logarithm of the radon measurement in house i (individual-level), in county j (group-level). On both levels we have predictors. The individual-level predictor, X_{ij}, is the floor in which the measurement was taken (0 for basement, 1 for first floor). The group-level predictor, Z_j, is the measurement of soil uranium that was available at county-level. ε_{ij} is the within-county variation, this includes measurement errors, natural variation in radon levels and variation between the different houses in the county beyond what is explained by the floor predictor. From Model (2), we have the group errors U_j, which in our example is the variation between counties beyond what is explained by the county-level predictor.

2.1 Model without Predictors

We look at the simplest random intercept model, the model without any predictors. This model contains a set of random groups with random variation within the groups. The model can be explained as a model where dependent variables is the sum of the mean, γ_{00}, a random part at group level U_{0j}, and a random part for the individual level, ε_{ij}. The random variables U_{0j} and ε_{ij} are assumed to have a mean of zero, to be mutually independent, and to have variances τ_0^2 and σ_ε^2 (Snijders and Bosker, 2012)

$$\begin{aligned} Y_{ij} &= \gamma_{00} + U_{0j} + \varepsilon_{ij}, \\ &\text{for } i = 1, ..., n_j. \\ &\text{and for } j = 0, ..., N. \end{aligned} \tag{3}$$

Model (3) does not have any predictors, but is important as it can describe the variability in the data between the independent-level and group-level. For example, low variability between the group-levels can suggest that we do not need to use multilevel modeling over the classical linear regression model.

The total variance in the observed values of Y_{ij} is the sum of the variances of $\text{var}(Y_{ij}) = \tau_0^2 + \sigma_\varepsilon^2$. Estimating these parameters in the model without predictors gives us the intraclass correlation coefficient (ICC). For the model without predictors the ICC explains the percentage of the total explained variance in the response accounted for by the belonging to a group:

$$\text{ICC} = \frac{\tau_0^2}{\tau_0^2 + \sigma_\varepsilon^2} \tag{4}$$

As we start adding predictors to the model, the understanding of ICC will change. The model without predictors contains just one fixed term and the variance at both levels. Calculating the ICC will give a proportion of the total variation at group-level and how similar the individuals within groups are compared to individuals in different groups. By adding variables to the fixed part at each level we can observe the change of the unexplained variation.

The ICC ranges from 0 to 1. When ICC is close to 0, group gives no information, but if the ICC is closer to 1, then there is no variance to explain at the individual level and individuals are alike inside each group. The ICC is used to see if we want to use a multilevel model or if the data is not grouped enough and we should return to the classical linear regression model. The ICC can be simply understood as the proportion of the variance that is explained by the grouping in the data.

Example 2 ICC from the Model without Predictors

We return to our example of home radon measurements in US homes. The response variable Y_{ij} is the logarithm of the radon measurement in the houses i (Level-1), within US counties j (Level-2). ha This gives us a nesting structure with homes within counties. Both levels have a predictor, which we will not take into account as we consider the model without predictors. We fit the random intercept model with no predictors

```
lmer(formula = y ~ 1 + (1 | county))
coef.est  coef.se
1.31      0.05

Error terms:
Groups    Name        Std.Dev.
county    (Intercept) 0.31
Residual              0.80
```

Fitting the model without predictors model gives us the parameters as shown in the code snippet above. The estimate τ_0 (group residual) is given as 0.31, while the estimate σ_ε (individual residual) is 0.80. This gives us the intraclass correlation coefficient $\frac{\tau_0^2}{\tau_0^2+\sigma_\varepsilon^2} = \frac{0.0961}{0.64+0.0961} \approx 0.13$. Approximately 10 percent of the variability lies at county-level. The ICC is not adequate for us to conclude upon the fact that we were right in choose the multilevel model.

2.2 Adding Predictors

The next step after making the model without predictors is to include predictors. Predictors are used to explain part of the variability of Y on both level-1 and level-2. Adding only one predictor for the individual-level, we get the model

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \tag{5}$$

$$\beta_{0j} = \gamma_{00} + U_{0j} \tag{6}$$

Here we assume that all residuals, U_{0j} and ε_{ij} have mean zero and are mutually independent given values X_{ij} of the predictor. Both residual are drawn from a normal distribution. The within-group residuals ε_{ij} is the same across the groups and shown as σ_ε^2. The U_{0j} is symbolized by τ_0^2 and can be seen as the residuals at group-level, or the group-effects that are unexplained by X. As residuals includes the variability of the dependent variable that is not modeled as a function of the predictors, the model has unexplained variability at two nested levels. We have four parameters, γ_{00}, β_1 and the two variance components τ_0^2 and σ_ε^2. The overall intercept γ_{00} is the intercept for the average group. β_1 is the regression coefficient or slope parameter, a unit increase in X gives an increase in the response of β_1.

The residual correlation between the Y values of two individuals in a group when controlling for X is

$$\rho(\mathrm{Y}|\mathrm{X}) = \frac{\tau_0^2}{\sigma_\varepsilon^2 + \tau_0^2}. \tag{7}$$

When we add predictors to the model, we see a change in the meaning of the ICC. We now interpret the ICC as a proportion of variance in the response controlling for the the predictors. We want to see if we by adding predictors at the different levels can see a change in the unexplained variation (Snijders and Bosker, 2012).

So for Equation (4), if the ICC = 0, this means $\mathrm{U}_{0\mathrm{j}} = 0$ for all of the groups j and grouping is not important for the response Y conditional on X. This means that we, in reality, could have used the classical linear regression model instead of the multilevel regression model.

Example 3 ICC when adding one Predictor

We add the predictor X of the individual-group, where the measurements were taken in the different houses (basement or first floor).

```
lmer(formula = y ~ x + (1 | county))
coef.est coef.se
(Intercept)  1.46     0.05
x           -0.69     0.07

Error terms:
Groups   Name        Std.Dev.
county   (Intercept) 0.33
Residual             0.76
```

The estimate τ_0 (group residual) is given as 0.1089, while the estimate σ_ε (individual residual) is 0.5776. This gives us the intraclass correlation coefficient $\frac{\tau_0^2}{\tau_0^2+\sigma_\varepsilon^2} = \frac{0.1089}{0.5776+0.1089} \approx 0.16$. We observe a slight increase in the ICC as the house-level predictor is added. As the predictor variable now accounts for a bigger proportion of the residual variation than the variation between the counties this makes sense.

We can now continue to add more predictors. Moving on we can add a predictor for the

group-level. The predictors for the individual-level is written as the values $X_{1i}, ..., X_{pi}$, and we chose to give the group-level the values $Z_1, ..., Z_q$. The model with two predictors, one for individual-level and one for county-level is given by Model (1).

Example 4 ICC when adding Predictors at both Levels

We add the county-level predictor Z_j, the measurement of soil uranium that is available at county-level. This gives the same equation as Equation (1). We use R to retrieve the model when adding both predictors.

```
lmer (y ~ x + u.full + (1| county))
coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.07
u.full         0.72        0.09

Error terms:
Groups   Name        Std.Dev.
county   (Intercept) 0.16
Residual             0.76
```

ICC $= \frac{0.0256}{0.0256+0.5576} \approx 0.044$. The ICC measure how much of the unexplained variation can be accounted for by the class we are in. When we add a county-level predictor, we are accounting for a bigger part of the variation between the different counties. The random intercept has less variation for this model than the ones without any county-level predictors, and so the ICC is also lower.

The county-level predictor leave us with an unchanged within-county variation, which makes sense as the group-level predictor can not explain variation within the counties. For small counties we are closer to complete pooling as a county would need at least 23 (1/ 0.044) observations to be drawn towards the no-pooling estimate rather than the complete-pooling estimate.

Figure 2 shows the multilevel regression line with uranium as a county-level predictor. The dashed lines show the previous regression line without the county-level predictor. For almost all counties, the addition of the predictor

does not seem to change the overall regression line, with the exception of two of the counties with fairly small sample-sizes, Aitkin and Koochiching, which have moved slightly towards the no-pooling estimate.

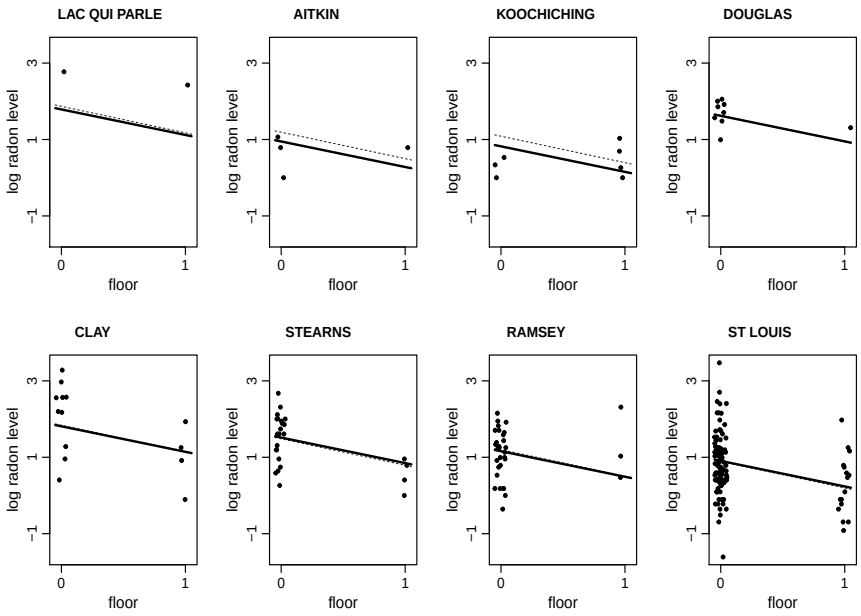


Figure 2: Multilevel regression lines fit to radon data, given for eight counties where we have included the county-level predictor, uranium. The dashed lines are the multilevel estimates without uranium as a predictor.

2.3 Estimating Parameters and the Random Group Effects

To be able to estimate the fixed effects and variance components of the model we refer to two different estimation methods (given that the residuals (ε_{ij} and U_{0j}) are normally distributed) maximum likelihood (ML) and residual maximum likelihood (REML) (Snijders and Bosker, 2012). The ML estimation includes the variance components (τ_0^2 and σ_ε^2) as well as the regression coefficients, so both the fixed effects part and random effects part in the likelihood function. The REML estimation includes only the variance components, so the parameters that sets the random effects part in the model. The REML method does however estimate the variance components as well as taking into account the loss of degrees of freedom derived from the estimation of the regression parameters. The ML method does not take into account the degrees of freedom lost when estimating the fixed

effects, giving the ML estimators for the variance components a negative incline (which we do not get with REML). Meaning that the ML estimates are biased when we have less observations. This difference can be important when the group sample-sizes are small. For groups with sample-size of more than thirty (given as a rule of thumb), the difference between the two methods will be insignificant (Snijders and Bosker, 2012).

Example 5 ML and REML

For the simple case of $y_i = \mu + \varepsilon_i$, the ML estimate of σ^2 is $\sum(y_i - \overline{y}_.)^2/n$, which is biased for σ_ε by a factor of (n-1)/n (Visscher et al., 2004). Here the ML method uses all the observations in our dataset whereas the REML method will use a likelihood function calculated from a transformed dataset, which includes only linear combinations of the responses that are unaffected by the intercept. The dimension of the data then n-1. For REML, the likelihood does not contain fixed effects and contain fewer terms giving us $\hat{\sigma}^2 = \frac{\sum(y_i - \overline{y}_.)^2}{n-1}$ (Duchateau et al., 1998).

The random intercept model is given by the parameters γ_{00}, β_1, τ_0^2 and σ_ε^2. Here the random group effect U_{0j} is not seen as a parameter, but rather a variable. For the groups with large group-sizes we have more information and so the uncertainty reduced and we see a larger effect of the estimates than on the groups with smaller sample-sizes. The influence of the group-size on the estimate is given by the ICC. If we look at estimating the mean intercept, γ_{00}, if the ICC equals to zero the different groups will have an affect on the estimated value of γ_{00}, so that it is equal to the group-size. Given that the ICC is one, every group will have the same affect, regardless of group-size. So, when the residual ICC lies between 0 and 1, groups with bigger group-sizes will have a larger affect (Snijders and Bosker, 2012).

The random group effects, the U_{0j}, which are seen as a variables, are not estimated as a part of the parameter estimation. But it can be interesting to 'estimate' them nevertheless. To estimate the group effects we use the *empirical Bayes estimation*. The *empirical Bayes estimation* estimates U_{0j} by using two different information sources: the data from a given group j and that the U_{0j} is a random variable with $U_{0j} \sim N(0, \tau_0^2)$ (Snijders and Bosker, 2012)

As γ_{00} is already an estimated parameter, estimating β_{0j} will be the equivalent to the

estimate U_{0j} and adding the γ_{00}. This means that estimating β_{0j} and U_{0j} is in theory the same, and we can estimate either if we have the estimate of γ_{00}. If we only estimate for group j, the β_{0j} would be estimated as the group mean

$$\hat{\beta}_{0j} = \overline{Y}_{.j} \tag{8}$$

For the whole, we would estimate β_{0j} by the sum of the general mean, γ_{00}. We estimate with the total mean

$$\hat{\gamma}_{00} = \overline{Y}_{..} = \sum_{j=1}^{N} \frac{n_j}{M} \overline{Y}_{.j}, \tag{9}$$

where n_j is the sample size of a given group j, and M is given as the total sample size, $M = \sum_j n_j$ It is given that the optimal estimates for β_{0j} is given by the weighted average of the two estimates mentioned above, Equation (8) and Equation (9):

$$\hat{\beta}^*_{0j} = \lambda_j \hat{\beta}_{0j} + (1 - \lambda_j)\hat{\gamma}_{00} \tag{10}$$

where λ_j is the reliability of the mean of group j.

$$\hat{\lambda}_j = \frac{\hat{\tau}^2_0}{\hat{\tau}^2_0 + \frac{\hat{\sigma}^2_\varepsilon}{n_j}}, \tag{11}$$

where $\hat{\tau}^2_0$ and $\hat{\sigma}^2_\varepsilon$ are found by REML. Equation (10) is called the empirical Bayes estimate for β_{0j}. The Equation (10) can be seen as the estimated group mean only pushed slightly to the overall mean γ_{00}, so that we have a shrinkage to the mean. When looking at Equation (11) it is quite obvious that the group j will be larger when n_j (sample size) is larger. This means that for larger groups the empirical Bayes estimate will almost be the same as the intercept estimated from data in group j, $\hat{\beta_{0j}}$.

Example 6 No-pooling vs. Partial-pooling

In Figure 3 we have the two plots where (a) shows the estimates for the county intercepts for the no-pooling analysis plotted against number of houses. The counties with fewer measurements have more variable estimates with higher standard errors. This illustrates the problem of classical regression as we deal with nested data, it makes us think that some counties are more extreme, just because they have small sample sizes. (b) shows the multilevel estimates for the county intercepts plotted against number of houses in the county. If we compare the left (no-pooling) and the right plot, we observe

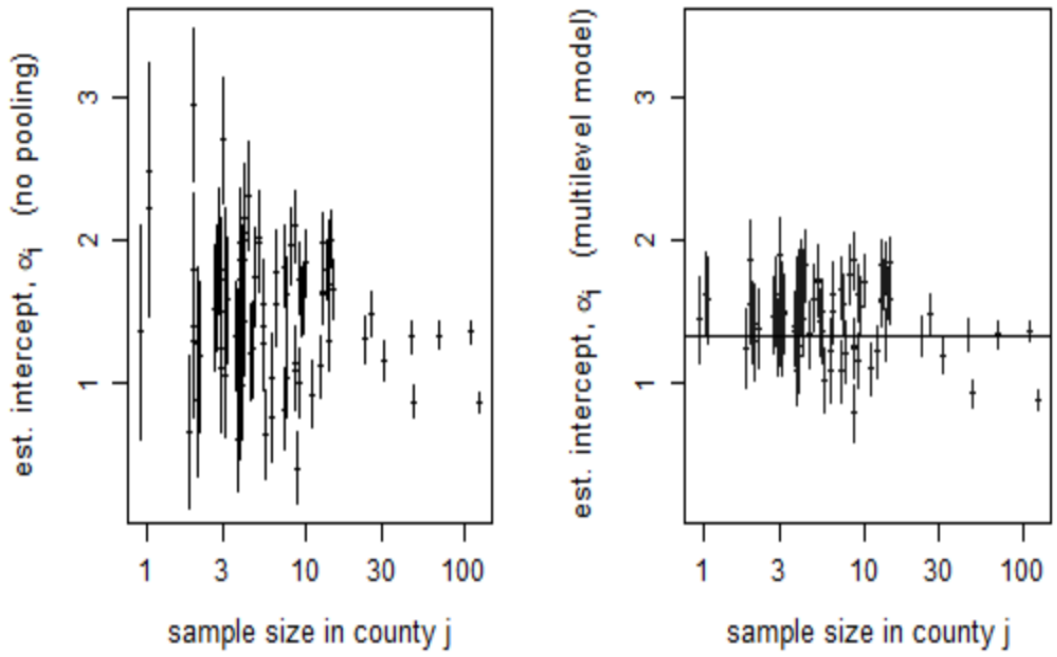


Figure 3: (a) The figure to the left. Estimates ± standard errors for the county intercepts for the no-pooling analysis of the radon data, plotted against number of houses in the county.

(b)The figure to the right. Multilevel (partial pooling) estimates ± standard error for the county intercepts for the radon data plotted against number of houses in the county. The horizontal line shows the complete-pooling estimate.

that the multilevel estimate is closer to the complete pooling estimate for counties with few houses, and closer to no-pooling for estimates for counties with many observations Gelman and Hill (2007). This is because of how we estimate with multilevel analysis. If we take a look at Equation (10) and (11) this shows us how the groups with large sample-sizes moves towards the no-pooling estimate and how the groups of small sample-sizes moves towards the overall mean.

If we look at Figure 1, we can recognize the same pattern here. For counties with small sample-sizes (few houses), the regression line moves towards the complete-pooling line. This is especially obvious for the county Lac Qui Parle, which only has two observations.

3 Random Slope Model

We have now looked at the simpler version of the multilevel model, where only the intercepts would vary between groups. Here the intercept was the only random part in the model. However, slopes can also be a random part in the multilevel model. For some groups, the predictors can have a large effect on the response, and for some groups, a small effect on the response. When this is the case, setting our slope to be random could give us a model which better fits our data. The next step is therefore to allow the regression coefficients to vary by group. By adding random slope to the random intercept model we are allowing individual-level relationships to vary across groups. The random slope, random intercept model is given by:

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}, \\ \text{for } i = 1, ..., n_j. \\ \text{and for } j = 0, ..., N. \end{gathered} \tag{12}$$

where β_{1j} is the regression coefficient which contains a random part allowing for the slopes to vary from different groups. This random part of the regression coefficient is the only added part to Model (3). ε_{ij} is normally distributed with mean zero and variance σ^2_ε, given that all residuals are independent of each other and identically distributed with variance constant across the groups. When we allow for the slope to be a random part in the model, it will makes sense for the intercept to vary as well. If the individual predictor varies by group, it will also makes sense that the intercept of the regression should too. In the next section we explain how we allow for the slope parameter to vary, as well as the intercept. Sometimes it is acceptable for the slope to vary without the intercept varying, such as when we want to control for some conditions in a study with multiple experiments while we let the treatments vary.

3.1 Random Intercept and Slope

As said, when allowing for the slope to vary, it is in most cases natural to let the intercept vary as well. For a random slope, random intercept model, we add a random term to the coefficient of X so that it can differ for all of the groups, and so the relationship between the response and X is different between groups. Both the intercept β_{0j} and the regression coefficients β_{1j} are now dependent by group. These are then split into the overall mean γ_{00} and γ_{10} and the random effects U_{0j} and U_{1j}. For model (12) the parts allowed to

vary are given by:

$$\beta_{0j} = \gamma_{00} + U_{0j} \tag{13}$$

$$\beta_{1j} = \gamma_{10} + U_{1j} \tag{14}$$

The pairs of random effects (U_{0j}, U_{1j}) are for different groups, independent and identically distributed and drawn from a bivariate normal distribution with mean zero and variance $\text{var}(U_{0j}) = \tau_{00} = \tau_0^2$, giving the variance in the effect of group j on the mean response variable. And $\text{var}(U_{1j}) = \tau_{11} = \tau_1^2$, being the variance in the effect of group j on the slope of the level-1 predictor. The covariance between level-1 intercepts and slopes is given as: $\text{cov}(U_{0j}, U_{1j}) = \tau_{01}$. The individual residuals ε_{ij} have mean zero and variance σ_ε^2. The two group effects U_{0j} and U_{1j} will usually not be independent, but correlated. We have:

$$\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \sim N(0, \Omega_u), \qquad \Omega_u = \begin{bmatrix} \tau_0^2 & \\ \tau_{01} & \tau_1^2 \end{bmatrix}$$

U_{1j} is the added random part for slope, and is the given as the difference between the slope of a group j and the slope of the overall line. The γ_{10} is given as the mean regression coefficient. We put in the substitutions and get the model

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + U_{0j} + U_{1j}X_{ij} + \varepsilon_{ij} \tag{15}$$

The first part of Model (15), $\gamma_{00} + \gamma_{10}X_{ij}$, is now the fixed part of the model, γ_{10} is the average regression coefficient while γ_{00} is the average intercept. The random part of the model is $U_{0j} + U_{1j}X_{ij} + \varepsilon_{ij}$, here $U_{1j}X_{ij}$ is the random interaction between group and predictor X. Model (15) is a model where the groups are represented by two random effects, intercept and slope. The two group effects will be correlated and not independent, as seen from the covariance matrix Ω_u. For the different groups of the model, the pairs of random effects U_{0j}, U_{1j}, are independent and identically distributed, and independent of the individual-level residual, ε_{ij} (Snijders and Bosker, 2012).

Random slopes can be understood as interactions between an individual-level predictor and group indicators. The intercepts can be interpreted easier if the predictor is centered, which can lead to lower correlation. It is not entirely wrong having a high correlation between the intercepts and slopes, but the estimated intercepts are more difficult to understand. It can therefore be favorable to remove the mean value of the continuous x before adding it in the regression. So for the model $Y_{ij} = \beta_{0j} + \beta_{1j}x_ij + \varepsilon_{ij}$, we get

that $x_{ij} = z_{ij} - \bar{z}$. This can lead to the correlation becoming closer to zero. Centering the X will not necessarily delete the correlation between the intercept and slope, but the remaining correlation can be more understandable and easier to explain.

3.2 Adding Predictors

We can expand the random intercept, random slope model by adding predictors. In Example 7 we look at the random intercept, random slope model for the radon data with only the individual-level predictor. In Example 8 we add the group-level predictor.

Example 7 Random intercept and slope with house-level predictor

We will illustrate Model (11) with the home radon example. This is an easy example as we only have one individual-level predictor, x. So for the random slope, random intercept model we include the house-level predictor x (floor measurement), but without the county-level predictor of uranium.

```
                lmer(formula = y ~ x + (1 + x | county))
            coef.est coef.se
(Intercept)  1.46     0.05
x           -0.68     0.09

Error terms:
 Groups   Name        Std.Dev. Corr
 county   (Intercept) 0.35
          x           0.34     -0.34
 Residual             0.75
```

For this model, the unexplained within-county variation has the estimated standard deviation of $\sigma_\varepsilon = 0.75$ and the estimated standard deviation of the county intercepts is $\tau_0 = 0.35$. THe estimated standard deviation of the county slopes is $\tau_1 = 0.34$ and estimated correlation between intercepts and slopes are -0.34.

We look at the average group coefficient and the estimated errors for each county. First we get the fixed effects by typing fixef(M3) (the estimated average coefficients.)

```
        fixef (M3)
        (Intercept)          x
          1.46               -0.68
```

Then to see the random effects (the estimated group-level errors):

```
        ranef (M3)
    (Intercept)            x
1  -0.32                0.14
2  -0.53               -0.09
3   0.009               0.012
4   0.07               -0.07
        .   .   .
85 -0.083               0.027
```

We get the estimated intercept, β_{0j}, and slope, β_{1j}, for each county by adding the errors to γ_{00} and γ_{10}. We are given that the estimated regression line for county 1 is (1.46 - 0.32) + (-0.68 + 0.14)X = 1.14 - 0.54X. The group-level model for (β_{0j}, β_{1j}) allows for partial-pooling in the estimated intercepts and slopes. Figure 4 shows the results as the estimated lines for the radon data in eighth different counties.

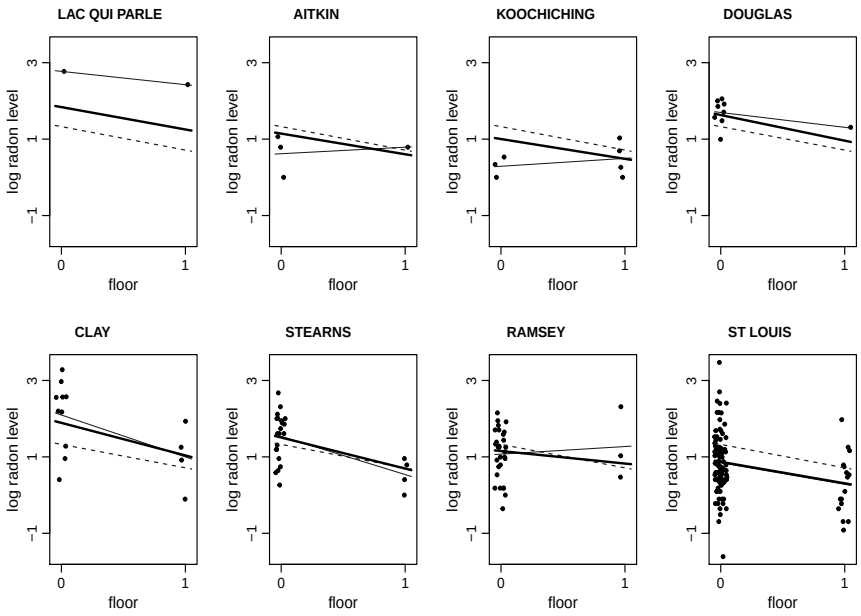


Figure 4: Multilevel regression for random slope and random intercept is seen as the thick, solid black line, no-pooling as the thin solid line and light coloured dashed line is the complete pooling regression.

Figure 4 shows the results of the random intercept, random slope estimated lines for the radon data for the eight different counties. We can compare this to figure 1, where

only the intercept is allowed to vary. We clearly observe a lot of pooling when the slope coefficient is estimated in the multilevel model, this is because the slope coefficient in each county are close to the complete-pooling estimate. At the same time we observe that for two of the counties with a large sample-size, Stearns and Ramsay, that the slope coefficient deviate from the common estimate. The high proportion of pooling for the slope coefficient explains why the estimates in Figure 4 are fairly similar to the ones in Figure 1.

Example 8 Random Intercept and Slope adding the Group-level Predictor

We can expand the in Example 7 model by adding a group-level predictor, the county-level predictor, soil uranium.

```
M4 <- lmer (y ~ x + u.full + x:u.full + (1 + x | county))
            coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.08
u.full       0.81     0.09
x:u.full    -0.42     0.23

Error terms:
 Groups   Name        Std.Dev. Corr
 county   (Intercept) 0.12
          x           0.31     0.41
 Residual             0.75
```

The estimates γ_{00}, U_{0j}, γ_{10} and U_{1j} are the coefficients for the intercept, x, u.full and x:u.full in the regression. The interaction corresponds to letting uranium be a predictor in the regression for slopes.

When adding the group-level predictors we can see that it reduces the group-level variation to 0.12, one third of the variation given in Example 7. This is because the group-level estimate induces stronger pooling. We can therefore make the assumption that this model would be better than the one portrayed in Example 7.

We can combine the average coefficients with the county-level errors to compute the intercepts and slopes as in Example 5.

If we were to expand the model we include more variables that have random effects, and more variables to explain the random effects. Imagine that we have p number of level-1

predictors $X_1, \ldots, X_p$ and q level-2 predictors $Z_1, \ldots, Z_p$. For individuals the model is a regression model with p variables:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \cdots + \beta_{pj}x_{pij} + \varepsilon_{ij} \tag{16}$$

The regression coefficients β_{0j} to β_{pj} is explained by the between-group model, which is a q- variable regression model for the group-dependent coefficient β_{hj}:

$$\beta_{hj} = \gamma_{h0} + \gamma_{h1}z_{1j} + \cdots + \gamma_{hq}z_{qj} + U_{hj} \tag{17}$$

Substituting and rearranging Model (16) and Model (17) gives us:

$$\begin{aligned} Y_{ij} = \gamma_{h0} + \textstyle\sum_{h=1}^{p} \gamma_{h0}x_{hij} + \sum_{k=1}^{q} \gamma_{0k}x_{kj} + \\ \textstyle\sum_{k=1}^{q}\sum_{h=1}^{p} \gamma_{zk}z_{kj}x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj}x_{hij} + \varepsilon_{ij} \end{aligned} \tag{18}$$

(Snijders and Bosker, 2012)

3.3 Estimating Parameters of Random Slope

What we mention in Section 2.3 regarding estimation of random intercept parameters, can to a certain degree be applied to the more complicated model of random slope as well. The random intercept, random slope model contain the parameters γ_{00},γ_{10}, β_{0j}, β_{1j}, τ_0^2, τ_1^2, τ_{01}^2 and σ_ε^2. The random effects are given by U_{0j} and U_{1j}, which we define as variables in the estimation. The variance parameters can, as we do for the random intercept model, be estimated with the ML or the REML method. The REML method is preferable as it produces less biased estimates for the parameters in the random part. When using lmer() method in R, the models are fit with REML. The random group effects (U_{0j}, U_{1j}) can be 'estimated' by the *empirical Bayes method*, and works similarly mentioned in section 2.3. (Snijders and Bosker, 2012).

As we move on from the random intercept model, we only have two new parameters to estimate, we've got the parameters τ_1^2, τ_{01}^2. It is assumed that the level-2 residuals U_{0j} and $U_{1j}x_{ij}$ together with the level-1 residual ε_{ij}, have means zero given the values of the level-1 predictor variables X. So that γ_{01} is the average regression coefficient just like γ_{00} is the average intercept. For random intercept, random slope, $\gamma_{00} + \gamma_{01}x_{ij}$ is referred to as the fixed part of the Model (15).

These random parameter estimates can not be interpreted separately, but have to be

done together due to correlation. At level-2 we got the random part U_{0j} and $U_{1j}x_{ij}$, where the term $U_{1j}x_{ij}$ is the random interaction between group and the level-1 predictor X. The model tells ut that the groups are defined by two random effects, the intercept and slope. For the level-2 variance we have:

$$\begin{aligned} \text{Var}(U_{0j} + U_{1j}x_{ij}) &= \text{var}(U_{0j}) + 2\text{cov}(U_{0j}, U_{1j}x_{ij}) + \text{var}(U_{1j}x_{ij}) \\ &= \tau_0^2 + 2\tau_{01}^2 x_{ij} + \tau_1^2 x_{ij}^2 + \sigma_\varepsilon^2 \end{aligned}$$

Where for we for two different individuals i and k in the same group, with $i \neq k$:

$$\text{cov}(Y_{ij}, Y_{kj}|x_{ij}, x_{kj}) = \tau_0^2 + \tau_{01}^2(x_{ij} + x_{kj}) + \tau_1^2(x_{ij}x_{kj})$$

4 Model Checking

Model checking for multilevel models is more complicated than model checking for classical linear regression models, as we have to take into account the different nested levels. When we have chosen the model we want to use to fit our data, we would like to compare fits with other models. Several assumptions are made when using the multilevel model, such as: residuals being normally distributed, random effects being normally distributed and uncorrelated etc. Ultimately we would like to compare fits with other models and evaluate plausibility of assumptions.

4.1 AIC and DIC

Given a set of nested models, AIC (Akaike information criterion) is used to quantify the fit of the different models. AIC is given as a number which we can use to compare models and choose the one assumed to be the best fit. For this AIC uses log-likelihood evaluated at the maximum likelihood estimates of the parameters, where the penalty is the number of parameters. We want the model with the lowest AIC value. For nested data however, it can be difficult the interpret the number of parameters. The AIC is defined as:

$$\text{AIC} = -2 \times \log - \text{likelihood} + 2 \times \text{number of parameters}$$

where -2 times the log-likelihood is -2 times the logarithm of the likelihood of the data given the estimated model parameters (Gelman and Hill, 2007). In classical regression a new model is estimated to reduce out-of-sample prediction error if the AIC decreases (Gelman and Hill, 2007)

For nested data it can be difficult to define a given number of parameters, as we in multilevel models use different amounts of pooling. For a complete-pooling model one parameter will equal one group of J parameters, while we with no-pooling a parameter will be equivalent to the J independent parameters. For the partial pooling we will get something in between complete pooling and no-pooling (Gelman and Hill, 2007).

For example, for the random intercept radon models, we have that the coefficients for the 85 county indicators represent less than about 85 independent parameters. For counties with small sample sizes the group-level regression explains much of the variation in the intercepts, so that in the multilevel model are not estimated independently. When the

model is improved and the group-level variance lowers, so will the effective number of independent variables (Gelman and Hill, 2007).

There are also other ways of performing model checking for nested data, for example with the use of deviance information criterion (DIC). This is often used in Bayesian statistics, which is out of the scope of this thesis. The idea, however, is that where we with AIC have a penalty for the number of effective parameters. For DIC the models also penalize by the value of the mean deviance, where the mean deviance is given as the –2× log-likelihood. The measure of out-of-sample predictive error, DIC, is given as

$$\text{DIC} = \text{mean deviance} + 2 \times \text{number of parameters.}$$

To find deviance we have used Bugs, as seen in the examples used in Gelman and Hill (2007) deviance is the $-2 \times \log - \text{likelihood}$.

Example 9 DIC: Comparing the fit of the models

We illustrate the use of DIC by comparing the fit of the models fit from our radon data.

Model	DIC
Random intercept without predictors	2251
Random intercept w/ individual-level predictor	2156
Random intercept w/ individual- and group-level predictor	2111
Random intercept, random slope without predictors	2154
Random intercept, random slope with predictors	2106

The model fit improves for the random intercept model as we add predictors. When the models get more complicated, the mean deviance decreases. As we expected with more structure we can fit the data better. The best model is given as the random intercept, random slope with predictors, where we allow for both the intercept and the slope to vary. This model will do best in predicting new houses (Gelman and Hill, 2007).

4.2 R^2

Different information criterias have often been used when comparing multilevel models. AIC and DIC are commonly used for multilevel models. These are used to select the best model or the model which is better than the others. There are however some limitations to the use when comparing AIC and DIC to R^2. While the information criteria gives us an estimate of the relative fit of different models, they will not give us the absolute model fit. Furthermore, AIC or DIC will not give any information about the variance explained by the model (Nakagawa and Schielzeth, 2013).

Snijders and Bosker (2012) use the definition of R^2 where the explained proportion of variance is given as the proportional reduction of prediction error, at both levels of the multilevel model. At level-1 we take a look at the prediction of the response variable Y_{ij}, where we have the level-one unit of i within a level-two group j. If the values of the predictors X_{ij} are not known, then the best predictor for the response variable is its expectation, so that the mean squared prediction error is $\text{var}(Y_{ij})$. If the predictor values are known, the best linear predictor for Y_{ij} is the regression value $\hat{Y}_{ij}$ so the mean squared prediction error is given as the sum of the residual variances at both levels. :

$$\text{var}(Y_{ij} - \hat{Y}_{ij}) = \sigma_e^2 + \tau_0^2. \tag{19}$$

the definition of the residual variances are the same as for Model (12). The level-1 explained proportion of variance is defined as the proportional reduction in mean squared prediction error:

$$R_1^2 = 1 - \frac{\text{var}(Y_{ij} - \hat{Y}_{ij})}{\text{var}(Y_{ij})} = \frac{\sigma_e^2 + \tau_0^2}{\text{var}(Y_{ij})} \tag{20}$$

For a sequence of nested models, the contribution to the estimated value of R_1^2 when adding predictors can be considered to be the contribution of the predictors to the explained variance at level 1 (Snijders and Bosker, 2012).

Example 10 Estimating the Level-1 Explained Variance

The easiest way to estimate R_1^2 is to consider $\hat{\sigma}_\varepsilon^2 + \hat{\tau}_0^2$ for the model with no predictors and for the random intercept model. From the radon example we have:

1. $Y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}, \qquad \sigma^2_\varepsilon = 0.0961, \;\; \tau^2_0 = 0.64$

2. $Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 Z_j + \varepsilon_{ij} \qquad \sigma^2_\varepsilon = 0.0256, \;\; \tau^2_0 = 0.5776$

$$R^2_1 = 1 - \frac{(0.0256 + 0.5776)}{(0.0961 + 0.64)} = 1 - 0.81945387854 \sim 0.181$$

The level-2 explained proportion of variance can be defined as the proportional reduction in mean squared prediction error, for the prediction of $\bar{Y}_{.j}$ for a group j, where $\bar{Y}_{.j}$ is the group mean. If the values of the predictors are unknown, then the best predictor $\bar{Y}_{.j}$ is the expectation, which is the mean squared prediction error $\mathrm{var}(\bar{Y}_{.j})$. Is the predictor values are known, then the best predictor of $\bar{Y}_{.j}$ is the regression value $\bar{Y}'_{.j}$, the associated mean square prediction error is

$$\mathrm{var}(\bar{Y}_{.j} - \bar{Y}'_{.j}) = \frac{\sigma^2_e}{n} + \tau^2_0. \tag{21}$$

where n is the number of level-1 units on which the average is based. The level-two explained proportion of variance is now defined as the proportional reduction in mean squared prediction error for $\mathrm{var}(\bar{Y}_{.j})$:

$$R^2_2 = 1 - \frac{\mathrm{var}(\bar{Y}_{.j} - \bar{Y}'_{.j})}{\mathrm{var}(\bar{Y}_{.j})} = \frac{\sigma^2_e + \tau^2_0}{\mathrm{var}(Y_{ij})} \tag{22}$$

(Snijders and Bosker, 2012)

Example 11 Estimating the level-2 explained variance

In the radon example we have 919 observations from 85 counties, we give n = 10 which is the representative value of the group size. We use the same values as in Example 11. For level-2 we then have:

0.0256 / 10 + 0.5776 = 0.58016

0.0961 / 10 + 0.64 = 0.64961

$R^2_2 = 1 - \frac{0.58016}{0.64961} \sim 0.1069$

For the case of varying group sizes one possibility could have been to use the harmonic mean, defined by $N/\sum_j (1/n_j)$ (Snijders and Bosker, 2012).

5 Discussion

Multilevel modeling makes it possible to describe variation at the different levels of the model. This is a consequence of the multilevel models ability to borrow strength from other groups by shrinkage, or partial pooling. However, a problems frequently returned to when dealing with multilevel models is uncertainty regarding the variability of the model. As we deal with a lot of parameters this adds a challenge to the interpretation. For our radon example we have 85 county-level coefficients for the random intercept model and 170 if we allow for the slope to vary (Gelman and Hill, 2007).

We mention the ICC for the random intercept models, as it can be useful to understand the relationship between groups. ICC is given as the correlation between two observations within the same group, so the higher the correlation within the group the lower the variability is within the cluster which means higher in between group variability. When the ICC is high this indicates that we were correct in choosing the random intercept model.

R squared is useful in the sense that we are able to understand how much of the variance is explained at the different levels in our model. R squared, which is fairly easily explained for a classical linear regression model have become increasingly difficult to calculate, as we now have to take into account the two levels. The ways of calculating R squared for multilevel models are many, and not all methods are agreed upon by different statisticians. For the radon data example we are left with the explained proportion of variance in level 1 being 0.18 and 0.10 for Level 2. Despite the checks, a lot of the total variation in radon levels across households in Minnesota remains unexplained with the different models. The multilevel analysis confirms that the presence of a basement gives an increased risk of higher radon levels in the household. Households in counties with high levels of uranium are also associated with higher radon levels. This might not be a shock to the reader, however, this is not necessarily the goal of the multilevel analysis. We could be looking for better estimates of the individual county, in particular the ones with smaller sample sizes. As we see from the model checking section, the DIC indicates that adding county-level predictor uranium improves the model. Differences in uranium measurements in counties helps explain not only the county difference in levels of radon, but the effect basements have on the radon level.

By using multilevel analysis we were able to get a number on the importance of variation within groups and show how multilevel analysis handles pooling. Random effect drags groups of large sample-sizes closer to no-pooling, while groups of small sample-sizes are dragged towards complete-pooling. When we use multilevel modeling, pooling happens automatically and we are able to rely more on the results as it is not something we have had to steer ourselves.

As this text is a short introduction to multilevel models, there are certain topics I have chosen not to get into. This includes generalized linear models and Bayesian analysis.

References

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 2017. doi: 10.18637/jss.v076.i01.

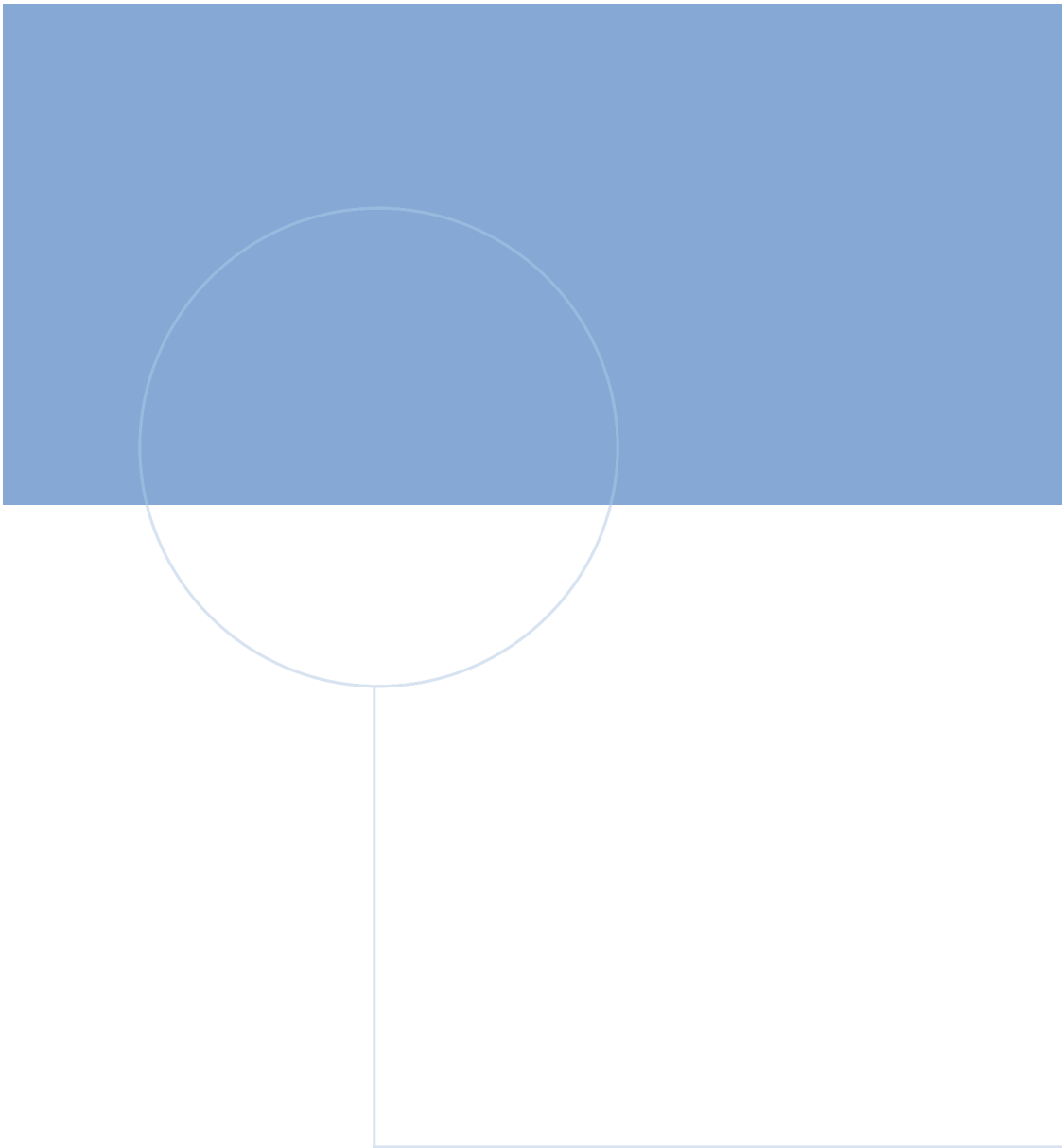
L. Duchateau, P. Janssen, and J. Rowlands. *Linear mixed models. An introduction with applications in veterinary research.* ILRI (International Livestock Research Institute, Nairobi, Kenya, 1998.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, United Kingdom, 2007.

S. Nakagawa and H. Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4:133–142, 2013. doi: 10.1111/j.2041-210x.2012.00261.x.

T. A. Snijders and R. J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage Publications, London ,United Kingdom, 2012.

P. M. Visscher, B. Benyamin, and I. White. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Cambridge University Press*, page 670–674, 2004. doi: 10.1375/1369052042663742.



NTNU

Norwegian University of
Science and Technology