**Category**

* [Advanced Modeling](https://datascienceplus.com/category/advanced-modeling/)

**Tags**

* [Linear Mixed Model](https://datascienceplus.com/tag/linear-mixed-model/)
* [R Programming](https://datascienceplus.com/tag/rstats/)

Regression analysis is one of the most demanding machine learning methods in 2019. One group of regression analysis for measuring hierarchical effects is Multilevel Models. This method is well suited for spatial differences between groups in the dataset. In this article, you learn how to do Multilevel Modelling in R.

**Multilevel models**

Many kinds of data, including observational data collected in the social sciences, health sciences and econometrics, have a hierarchical or clustered structure. For example, children with the same parents tend to be more alike in their physical and mental characteristics than individuals chosen at random from the population at large. Individuals may be further nested within geographical areas or institutions such as schools or employers. Multilevel data structures also arise in longitudinal studies where an individual’s responses over time are correlated with each other.

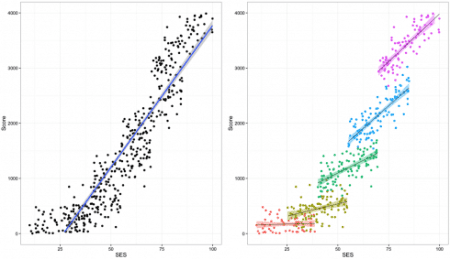
Multilevel models recognize the existence of such data hierarchies by allowing for residual components at each level in the hierarchy. For example, a two-level model which allows for the grouping of child outcomes within schools would include residuals at the child and school level. Thus the residual variance is partitioned into a between-school component (the variance of the school-level residuals) and a within-school component (the variance of the child-level residuals). The school residuals, often called ‘school effects’, represent unobserved school characteristics that affect child outcomes. It is these unobserved variables which lead to a correlation between outcomes for children from the same school.

Multilevel models can also be fitted to non-hierarchical structures. For instance, children might be nested within a cross-classification of neighborhoods of residence and schools.

**Why use multilevel models**

Multilevel models can be used to correct inferences: Traditional multiple regression techniques treat the units of analysis as independent observations. One consequence of failing to recognize hierarchical structures is that standard errors of regression coefficients will be underestimated, leading to an overstatement of statistical significance. Standard errors for the coefficients of higher-level predictor variables will be the most affected by ignoring grouping.

Substantive interest in group effects: In many situations a key research question concerns the extent of grouping in individual outcomes, and the identification of ‘outlying’ groups. In evaluations of school performance, for example, interest centres on obtaining ‘value-added’ school effects on pupil attainment. Such effects correspond to school-level residuals in a multilevel model which adjusts for prior attainment. It is possible to plot multilevel models with the below graph:

[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2019/07/MLM.png?ssl=1)

Here we can see that we have estimated a traditional regression model in the first graph and a multilevel model in the second graph.

Estimating group effects simultaneously with the effects of group-level predictors: An alternative way to allow for group effects is to include dummy variables for groups in a traditional (ordinary least squares) regression model. Such a model is called an analysis of variance or fixed-effects model. In many cases, there will be predictors defined at the group level, eg type of school (mixed vs. single-sex). In a fixed-effects model, the effects of group-level predictors are confounded with the effects of the group dummies, ie it is not possible to separate out effects due to observed and unobserved group characteristics. In a multilevel (random effects) model, the effects of both types of a variable can be estimated.  
Inference to a population of groups: In a multilevel model the groups in the sample are treated as a random sample from a population of groups. Using a fixed-effects model, inferences cannot be made beyond the groups in the sample.

**Methodology in Multilevel Models: Random effects**

In order to run a multilevel model, you have to be clear about what is fixed and what is random in your model. In the Random effects model, individual-specific effects are uncorrelated with the independent variables. Therefore the following are specific for the mode: Dummies act as an error term. Data consist of a hierarchy of populations whose differences relate to that hierarchy. We also estimate variance components for groups and error, assuming the same intercept and slopes. The difference among groups (or time periods) lies in their variance of the error term, not in their intercepts.

Consider the following model:

$$ math\_{ij}=β\_{0j}+β\_1\*homework\_{ij}+ε\_{ij} $$

Here, we have specified that the intercept varies by group (which is, in this case, the school). As such, we need to include another model for the random intercepts, but without a random slope.

$$ β\_{0j}=γ\_{00}+u\_{0j} $$  
$$ β\_{1j}=γ\_{10} $$

The notation here mirrors the notation of random effects. If we combine the formulas, we get the following:

$$ math\_{ij}=γ\_{00}+γ\_{10}\*(homework\_{ij})+u\_{0j}+ε\_{ij} $$

This is a random intercepts model, with fixed slopes.

**Methodology in Multilevel Models: Fixed effects**

In the Fixed effects model, individual-specific effect is correlated with the independent variables. Therefore the following are specific for the mode: Dummies are considered part of the intercept. We, therefore, examine group differences in intercepts. We also assume that the same slopes and constant variance across entities or subjects.  
The “Fixed Effects” section contains, labeled, the values of γ00 and γ10. Let’s say that we now want to include a random slope for β1. We would adjust the equation as follows:

$$ math\_{ij}=β\_{0j}+β\_1homework\_{ij}+ε\_{ij} $$

And we add the following on top of our earlier β0j equation:

$$ β\_{1j}=γ\_{10}+u\_{1j} $$

Yielding the following combined equation:

$$ math\_{ij}=γ\_{00}+γ\_{10}\*homework\_{ij}+u\_{0j}+u\_{1j}\*homework\_{ij}+ε\_{ij} $$

**Multilevel models in R: Random intercept model**

First, let’s load relevant packages and read the [data](https://github.com/knl84/Imm10-data).

# Import xlsx and Load dataset into R

library(readxl)

mlmdata <- read\_xlsx(file="/imm10.xlsx",sep = ",", header=TRUE) # read sheet

These are data containing, at the student level, information about math scores, socioeconomic status, sex, race, and other student characteristics. School-level characteristics include mean socioeconomic status, urbanicity, teacher/student ratios, and other characteristics.

Next we load relevant packages into R and run the statistical model:

# Load package

library(lme4)

Loading required package: Matrix

# Run random intercept model

model <- lmer(math ~ homework + (1 | schid), data=mlmdata)

Let’s explain this: You recognize the normal notation, math ~ homework. So what is $$ (1 | schid) $$This is the random effect? In other words, everything to the left of the | indicates the effects that should be random, and the variable to the right of the | is the grouping variable across which the effects should vary. What is the “1”? It’s the way we refer to the intercept. It’s not technically necessary in some cases, but it’s safe to just always include it in your random effects specification, I think.

As per usual, we can use the summary() command to see the details of our work.

# View summary of results

summary(model)

*Linear mixed model fit by REML ['lmerMod']*

*Formula: math ~ homework + (1 | schid)*

*Data: mlmdata*

*REML criterion at convergence: 1839.9*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-2.6060 -0.6872 -0.0244 0.5983 3.3770*

*Random effects:*

*Groups Name Variance Std.Dev.*

*schid (Intercept) 25.22 5.022*

*Residual 64.52 8.033*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error t value*

*(Intercept) 44.982 1.803 24.949*

*homework 2.207 0.379 5.823*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.371*

Note that R uses restricted maximum likelihood to fit the model. If you turn this off with the REML=FALSE option, it will use the optimization of the log-likelihood instead to fit the model, which aligns with some other software programs like Stata. In the “Random Effects” section of the results, you can see the random intercept. Using the notation from Raudenbush and Bryk (2006) again, 5.02 is the value of τ00, which is the standard deviation of u0j. The “Residual” standard deviation refers to σ.

**Multilevel Models in R: Random intercept and slope model**

# Run random intercept and slope model

model <- lmer(math ~ homework + (1 + homework | schid), data=mlmdata)

summary(model)

*Linear mixed model fit by REML ['lmerMod']*

*Formula: math ~ homework + (1 + homework | schid)*

*Data: mlmdata*

*REML criterion at convergence: 1764*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-2.5110 -0.5357 0.0175 0.6121 2.5708*

*Random effects:*

*Groups Name Variance Std.Dev. Corr*

*schid (Intercept) 69.30 8.325*

*homework 22.45 4.738 -0.81*

*Residual 43.07 6.563*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error t value*

*(Intercept) 44.771 2.744 16.318*

*homework 2.040 1.554 1.313*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.804*

Notice that there is now a standard deviation on “homework” in the “Random Effects” section of the output. This is the value of τ11, or the standard deviation of u1j. There is also a correlation between u0j and u1j of -0.81. This is τ01. What if we wanted to have a fixed intercept, and a random slope? Unfortunately, it’s not as easy as just taking out the “1” in the formula. If there is something on the left side of the |, and there isn’t a “1”, R will assume that you still want a random intercept. See below.

**Multilevel Models in R: Random slope model**

# Run random slope model

model <- lmer(math ~ homework + (homework | schid), data=mlmdata)

summary(model)

*Linear mixed model fit by REML ['lmerMod']*

*Formula: math ~ homework + (homework | schid)*

*Data: mlmdata*

*REML criterion at convergence: 1764*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-2.5110 -0.5357 0.0175 0.6121 2.5708*

*Random effects:*

*Groups Name Variance Std.Dev. Corr*

*schid (Intercept) 69.30 8.325*

*homework 22.45 4.738 -0.81*

*Residual 43.07 6.563*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error t value*

*(Intercept) 44.771 2.744 16.318*

*homework 2.040 1.554 1.313*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.804*

As shown in the above table, it still included a random intercept. To keep the intercept fixed while keeping the random slope, replace the “1” with a “0”.

# Run random slope model

model <- lmer(math ~ homework + (0 + homework | schid), data=mlmdata)

summary(model)

*Linear mixed model fit by REML ['lmerMod']*

*Formula: math ~ homework + (0 + homework | schid)*

*Data: mlmdata*

*REML criterion at convergence: 1849.1*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-2.15440 -0.72307 0.02491 0.69159 2.34777*

*Random effects:*

*Groups Name Variance Std.Dev.*

*schid homework 5.249 2.291*

*Residual 67.316 8.205*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error t value*

*(Intercept) 46.0301 0.9121 50.47*

*homework 1.6352 0.8685 1.88*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.434*

One last thing: Remember the value of τ01 from the random intercept and slope model. Sometimes, you may choose to fix the correlations between the random effects to 0. This is a modeling choice, and it’s easy to implement. All you need to do is include the random effects in separate terms in the model. See below.

**Multilevel Models in R: Random intercept and slope model**

# Run random intercept and slope model

model <- lmer(math ~ homework + (1 | schid) + (0 + homework | schid), data=mlmdata)

summary(model)

*Linear mixed model fit by REML ['lmerMod']*

*Formula: math ~ homework + (1 | schid) + (0 + homework | schid)*

*Data: mlmdata*

*REML criterion at convergence: 1772.7*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-2.55162 -0.54081 0.00279 0.62340 2.62067*

*Random effects:*

*Groups Name Variance Std.Dev.*

*schid (Intercept) 63.35 7.959*

*schid.1 homework 20.03 4.475*

*Residual 43.27 6.578*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error t value*

*(Intercept) 44.810 2.633 17.016*

*homework 2.016 1.475 1.367*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.065*

Now, no correlations are reported for the random effects, because they were set to 0.

**Multilevel logistic models**

Remember how switching from ordinary least squares regression (using lm()) to logistic regression (using glm()) required a shift to a generalized linear model? Surprise – the same thing happens here. We shift to the glmer() function, which has the same construction as lmer(). To specify that you want a logistic regression model, use the family=binomial(link="logit") option.

# Run logistic model with random intercept and slope

model <- glmer(white ~ homework + (1 + homework | schid), data=mlmdata,

family=binomial(link="logit"))

summary(model)

*Generalized linear mixed model fit by maximum likelihood (Laplace*

*Approximation) [glmerMod]*

*Family: binomial ( logit )*

*Formula: white ~ homework + (1 + homework | schid)*

*Data: mlmdata*

*AIC BIC logLik deviance df.resid*

*182.4 200.2 -86.2 172.4 255*

*Scaled residuals:*

*Min 1Q Median 3Q Max*

*-4.3373 -0.1184 0.1112 0.3421 3.8801*

*Random effects:*

*Groups Name Variance Std.Dev. Corr*

*schid (Intercept) 16.28733 4.0358*

*homework 0.04678 0.2163 -1.00*

*Number of obs: 260, groups: schid, 10*

*Fixed effects:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) 1.67362 1.47324 1.136 0.256*

*homework 0.04508 0.18433 0.244 0.807*

*Correlation of Fixed Effects:*

*(Intr)*

*homework -0.590*