Multilevel Modeling in R

Richard Blissett

2017-11-26

This review covers the *basics* of running a multilevel model in R. Very importantly, the data in this section were created by the wonderful people over at the UCLA Institute for Digital Research and Education (https://stats.idre.ucla.edu/) (IDRE). They also have great overviews of how to do various data analyses in R (and other languages), including multilevel modeling (https://stats.idre.ucla.edu/r/examples/mlm-imm/r-kreft-chp-3/). My only reasons for recreating overviews here is to align them with the tutorials you have gone through thus far and ensure their alignment with the curriculum at Seton Hall University. In addition, this overview is *not* for anyone learning multilevel modeling for the first time. For that, people should consult sources such as Raudenbush and Bryk (2002) (https://us.sagepub.com/en-us/nam/hierarchical-linear-models/book9230) or Gelman and Hill (2006) (http://www.stat.columbia.edu/~gelman/arm/). The latter resource is an especially useful resource for running multilevel models in R.

Multilevel linear models

First, let’s read in some data.

```
# Read in data
library(haven)
mlmdata <- read_dta("https://stats.idre.ucla.edu/stat/examples/imm/imm10.dta")
```

What command would you use to see the list of variables in these data?

Run:

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.

Just to practice, can you run some summary statistics on the data?

- % of students that are white:
- Average math scores:
- Total number of schools:

In order to run a multilevel model, you have to be clear about what is fixed and what is random in your model. Consider the following model.

$$math_{ij} = \beta_{0j} + \beta_1(homework_{ij}) + \varepsilon_{ij}$$

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.

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

The notation here mirrors the notation of Raudenbush and Bryk (2002). If we combine the formulas, we get the following.

$$math_{ij} = \gamma_{00} + \gamma_{10}(homework_{ij}) + u_{0j} + \varepsilon_{ij}$$

This is a random intercepts model, with fixed slopes.

To run a multilevel *linear* model, we use the `lmer()` function (“Linear Mixed Effects in R”) from the `lme4` package. The syntax will look very similar to the syntax from all of the regression functions we have used thus far.

```
# 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 τ_{00}, which is the standard deviation of u_{0j}. The “Residual” standard deviation refers to σ.

The “Fixed Effects” section contains, labeled, the values of γ_{00} and γ_10.

Let’s say that we now want to include a random slope for β_1. We would adjust the equation as follows:

$$math_{ij} = \beta_{0j} + \beta_1(homework_{ij}) + \varepsilon_{ij}$$

And we add the following on top of our earlier β_{0j} equation:

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Yielding the following combined equation:

$$math_{ij} = \gamma_{00} + \gamma_{10}(homework_{ij}) + u_{0j} + u_{1j}(homework_{ij}) + \varepsilon_{ij}$$

```
# 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 τ_{11}, or the standard deviation of u_{1j}. There is also a correlation between u_{0j} and u_{1j} of -0.81. This is τ_{01}.

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.

```
# 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
```

See? 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 τ_{01} 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.

```
# 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 - 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
```

Next steps

Got missing data? Let’s try out some multiple imputation: http://rpubs.com/rslbliss/r_multiple_imputation_ws (http://rpubs.com/rslbliss/r_multiple_imputation_ws)