IDS 702: Module 4.4

Multilevel/hierarchical linear models (illustration II)

Dr. Olanrewaju Michael Akande

The radon analysis cont'd

Variable	Description
radon	radon levels for each house
log_radon	log(radon)
state	state
floor	lowest living area of each house: 0 for basement, 1 for first floor
countyname	county names
countyID	ID for the county names (1-85)
fips	state + county fips code
uranium	county-level soil uranium
log_uranium	log(uranium)

INCLUDING GROUP-LEVEL PREDICTORS

- We should also control for uranium since radon occurs naturally as an indirect decay product of uranium.
- However, since each county has one single value for `uranium`, each house within that county has the exact same value.
- Turns out that including group-level predictors is quite straightforward in R, as long as the predictor is properly represented in the data as repeated values for all observations in the same group.
- One can ask the question: with 85 counties in the dataset, how are we able to fit a regression with 85 different intercepts for each county as well as a county-level coefficient for uranium?
- The simple answer is that we are actually using all the observations within each county (along with all observations from other counties in fact), when estimating each random intercept, but yes we only use 85 distinct values to estimate the effect of uranium.

The radon analysis: varying-intercepts

- Word of caution: be careful when including random slopes. You should really include them if you absolutely have to and if you have enough data to estimate them accurately.

- `lme4` in R uses the frequentist approach which is not fully reliable here as it uses an approximation for inference and it does not fully account for uncertainty in the estimated variance parameters. Personally, I prefer to use Bayesian models for multilevel regressions.

- If you want to fit a multilevel model for your final project, I would suggest taking a look at the `brms` package in R for a Bayesian approach.

- Let's use AIC to see if we can exclude the random slopes.

```
Model1 <- lmer(log_radon ~ floor + (floor | countyname), data = Radon)
Model2 <- lmer(log_radon ~ floor + (1 | countyname), data = Radon)
AIC(Model2); AIC(Model1) #same overall conclusions using BIC
```

```
## [1] 2179.305
```

```
## [1] 2180.325
```

- No real difference. We will exclude them going forward. You should be able to interpret the updated coefficients of the new model.

The radon analysis: including uranium

Turns out that it also often makes sense to use `log_uranium` instead of `uranium`.

```
Model3 <- lmer(log_radon ~ floor + log(uranium) + (1 | countyname), data = Radon) ; summary(Model3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_radon ~ floor + log(uranium) + (1 | countyname)
##    Data: Radon
##
## REML criterion at convergence: 2134.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9673 -0.6117  0.0274  0.6555  3.3848
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  countyname (Intercept) 0.02446  0.1564
##  Residual               0.57523  0.7584
## Number of obs: 919, groups:  countyname, 85
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)       1.46576    0.03794  38.633
## floorFirst Floor -0.66824    0.06880  -9.713
## log(uranium)      0.72027    0.09176   7.849
##
## Correlation of Fixed Effects:
##             (Intr) flrFrF
## florFrstFlr -0.357
## log(uranim)  0.145 -0.009
```

For any house in Minnesota with a basement as the lowest living area, every unit increase in log(uranium) increases radon levels by a multiplicative effect of $e^{0.72} = 2.05$.

How much data and how many groups?

- When J, that is the number of groups, is small, it is difficult to estimate the across-group variation.
- Multi-level modeling often adds little in such scenarios.
- However, it should not do any worse than including the grouping variable as a factor variable, and it can still be easier to interpret since we need not drop any level as baseline.
- Small sample sizes within the groups can be enough to fit a multilevel model when only the intercept is varying.
- With varying slopes, one can easily run into convergence issues.
- When groups do not have that many data points, the random intercepts and slopes may not be estimated accurately but the data within each group will still provide information that allows estimation of fixed effects and overall variance parameters.

Extra nested levels

- It is easy to envision applications where there might be more than one level of hierarchy.
- For example
  - students within schools within counties within states
  - patients within hospitals within states
  - voters within voting districts within states
- In those applications, it is straightforward to extend these ideas and create extra levels of hierarchy in the multi-level models.
- When that is the case, I once again prefer to rely on Bayesian methods to fit those models.

Non-nested models

- In other applications, there can be complicated grouping structures, where observations fall into two or more different non-nested grouping variables.
- For example
  - patients within J hospitals receiving K different treatments
  - students within J schools taking classes based on K different teaching techniques.
- Once again, it is straightforward to incorporate these within the context of multi-level models.

Non-nested models

- Suppose we want to fit a multi-level model with varying-intercepts by each grouping variable but with a fixed slope for one predictor, we would have

$$
\begin{aligned}
y_{ijk} &= (\beta_0 + \gamma_{0j} + \eta_{0k}) + \beta_1 x_{1ijk} + \epsilon_{ijk} \\
\gamma_{0j} &\sim N(0, \tau^2_{\gamma(0)}) \\
\eta_{0k} &\sim N(0, \tau^2_{\eta(0)}) \\
\epsilon_{ij} &\sim N(0, \sigma^2) \\
i &= 1, \ldots, n_{jk}; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K.
\end{aligned}
$$

- In R, we can fit the model above as follows:

```
M1 <- lmer(y ~ x + (1 | GroupVar1) + (1 | GroupVar2)) ; summary(M1)
```

- Adding more predictors is trivial.
- It is easy to add more group variables but it can be hard to fit the model without enough data points.

What's next?

Move on to the readings for the next module!