Multilevel Modeling — An Introduction

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multilevel Regression Modeling, 2009

Multilevel Modeling — An Introduction

Introduction

This lecture begins our detailed study of multilevel modeling procedures.

We concentrate in this lecture on an approach using R and the `lmer()` function.

Make sure that the `lme4` package is installed on your computer.

The Radon Study

One of the introductory examples in Gelman & Hill , and our first example of multilevel modeling, concerns the level of radon gas in houses in Minnesota.

Radon is a carcinogen estimated to cause several thousand lung cancer deaths per year in the U.S.

The Radon Study

The distribution of radon in American houses varies greatly. Some houses have dangerously high concentrations.

The EPA did a study of 80,000 houses throughout the country, in order to better understand the distribution of radon.

Two important predictors were available:

- Whether the measurement was taken in the basement, or the first floor, and
- The level of uranium in the county

Higher levels of uranium are expected to lead to higher radon levels, in general. And, in general, more radon will be measured in the basement than on the first floor.

The Radon Study

The distribution of radon in American houses varies greatly. Some houses have dangerously high concentrations.

The EPA did a study of 80,000 houses throughout the country, in order to better understand the distribution of radon.

Two important predictors were available:

- Whether the measurement was taken in the basement, or the first floor, and
- The level of uranium in the county

Higher levels of uranium are expected to lead to higher radon levels, in general. And, in general, more radon will be measured in the basement than on the first floor.

Hierarchical Data

The data are organized *hierarchically* in the radon study.

Houses are situated within 85 counties. Each house has a `radon` level that is the outcome variable in the study, and a binary `floor` indicator (0 for basement, 1 for first floor) which is a potential predictor.

Uranium levels are measured at the county level. There are 85 counties, and for each one a uranium background level is available.

We say that the level-1 data is at the house level, and the level-2 data is at the county level. Houses are grouped within counties.

Organizing Hierarchical Data

There are a number of ways to organize hierarchical data, and a number of different ways to write the same hierarchical model. One method breaks the data down by levels, and links the data through an intermediary variable.

This method offers some important advantages. It saves some space, and it emphasizes the hierarchical structure of the data.

Two Files for Two Levels

The level-1 file looks like this.

	county	radon	floor
1	1	2.2	1
2	1	2.2	0
3	1	2.9	0
4	1	1.0	0
5	2	3.1	0
6	2	2.5	0
7	2	1.5	0
.	.	.	.
.	.	.	.
917	84	5.0	0
918	85	3.7	0
919	85	2.9	0

Two Files for Two Levels

The level-2 file looks like this

	county	uranium
1	1	-0.689047595
2	2	-0.847312860
3	3	-0.113458774
.	.	.
.	.	.
85	85	0.355286981

A Single File for All Levels

An alternative, less efficient file structure puts all the data in the same file.

By necessity, some data are redundant.

The full data file looks like this:

	radon	floor	uranium	county
1	0.78845736	1	-0.689047595	1
2	0.78845736	0	-0.689047595	1
3	1.06471074	0	-0.689047595	1
4	0.00000000	0	-0.689047595	1
5	1.13140211	0	-0.847312860	2
6	0.91629073	0	-0.847312860	2
.	.	.	.	.
.	.	.	.	.
917	1.60943791	0	-0.090024275	84
918	1.30833282	0	0.355286981	85
919	1.06471074	0	0.355286981	85

"Old-Fashioned" Approaches

We have potential sources of variation at the county level, and at the house level. There are a number of potential approaches to analyzing such data that people have used prior to the popularization of multilevel modeling.

Two such approaches, discussed by Gelman & Hill , are

- *Complete Pooling.* Completely ignore the fact that the relationship between radon and uranium might vary across counties, and simply pool all the data. This model is

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

- *No Pooling.* Include county as a categorical predictor in the model, thereby adding 85 county indicators to the model.

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{2}$$

Fitting the Complete-Pooling Regression

First, we fit the complete-pooling model:

```
> radon.data ← read.table("radon.txt",header=TRUE)
> attach(radon.data)

> complete.pooling ← lm(radon ~ floor)
> display(complete.pooling)

lm(formula = radon ~ floor)
            coef.est coef.se
(Intercept)  1.33     0.03
floor       -0.61     0.07
---
n = 919, k = 2
residual sd = 0.82, R-Squared = 0.07
```

Fitting the No-Pooling Regression

```
> no.pooling <- lm(radon~floor + factor(county)-1)
> display(no.pooling)
```

```
lm(formula = radon ~ floor + factor(county) - 1)
                 coef.est coef.se
floor            -0.72     0.07
factor(county)1   0.84     0.38
factor(county)2   0.87     0.10
factor(county)3   1.53     0.44
.                 .        .
.                 .        .
factor(county)84  1.65     0.21
factor(county)85  1.19     0.53
---
n = 919, k = 86
residual sd = 0.76, R-Squared = 0.77
```

Basic 2-Level Models

At level 1, we have `floor` as a potential predictor of `radon` level.

We can think of the linear regression relating `floor` to `radon` in very simple terms.

The y-intercept is the average radon value at in the basement, i.e., when `floor = 0`.

The slope is the difference between average radon levels in the basement and first floor.

There are a number of ways we could model the situation.

Basic 2-Level Models

Our data are organized within county. Even in such a simple situation, there are numerous potential models for the relationship between radon level and floor.

- The slopes could vary across counties
- The intercepts could vary across counties
- Both the slopes and intercepts could vary

Gelman & Hill introduce a notation we can familiarize ourselves with, although it will take a little effort getting used to. Let's diagram these basic models and write them in the Gelman & Hill "full data" notation.

Basic 2-Level Models

Our data are organized within county. Even in such a simple situation, there are numerous potential models for the relationship between radon level and floor.

- The slopes could vary across counties
- The intercepts could vary across counties
- Both the slopes and intercepts could vary

Gelman & Hill introduce a notation we can familiarize ourselves with, although it will take a little effort getting used to. Let's diagram these basic models and write them in the Gelman & Hill "full data" notation.

Basic 2-Level Models

Our data are organized within county. Even in such a simple situation, there are numerous potential models for the relationship between radon level and floor.

- The slopes could vary across counties
- The intercepts could vary across counties
- Both the slopes and intercepts could vary

Gelman & Hill introduce a notation we can familiarize ourselves with, although it will take a little effort getting used to. Let's diagram these basic models and write them in the Gelman & Hill "full data" notation.

Varying Intercepts, No Predictor

One model allows the intercepts to vary across county, and uses no predictors. This model, which is formally equivalent to a one way random-effects ANOVA, can be written as

$$y_i = \alpha_{j[i]} + \epsilon_i \tag{3}$$

with

$$\epsilon_i \sim \mathrm{N}(0, \sigma_y^2) \tag{4}$$

and

$$\alpha_{j[i]} \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2) \tag{5}$$

In the above notation, "$j[i]$" means "the value of j assigned to the *ith* unit."

Varying Intercepts, No Predictor

```
> M0 <- lmer(radon ~ 1 + (1 | county))
> display(M0)

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

Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.31
 Residual             0.80
---
number of obs: 919, groups: county, 85
AIC = 2265.4, DIC = 2251
deviance = 2255.2
```

Varying Intercepts, Floor Predictor

The next model adds the `floor` predictor, and keeps varying intercepts. This model can be written as

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{6}$$

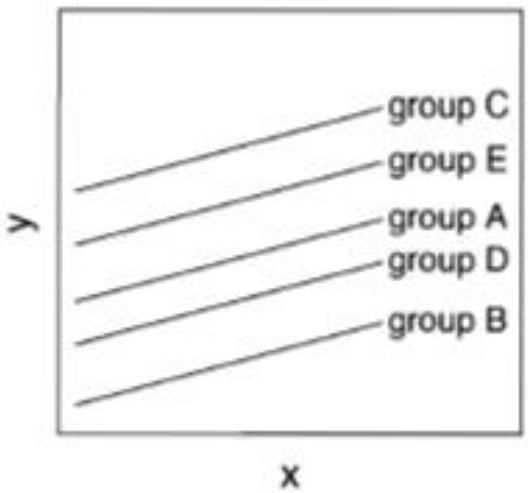
with

$$\alpha_{j[i]} \sim \mathrm{N}(\mu_\alpha, \sigma^2_\alpha) \tag{7}$$

This model looks much like the "no-pooling" model we looked at before, except that the earlier model used least squares estimation and essentially set each α to the value obtained by fitting regression within a county. Multilevel modeling uses a simultaneous estimation approach that is more sophisticated at dealing with large differences in sample size across counties.

Varying Intercepts, Floor Predictor

Here is a picture of the model with 5 counties:



Varying Intercepts, Floor Predictor

Here is how we fit this model using R.

```
> M1 ← lmer(radon ~ floor + (1 | county))
> display(M1)
```

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

Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.33
 Residual             0.76
---
number of obs: 919, groups: county, 85
AIC = 2179.3, DIC = 2156
deviance = 2163.7
```

Varying Intercepts, Floor Predictor

This model displays fixed and random effect results. To see more detail, we can use the `summary()` function.

```
> summary(M1)

Linear mixed model fit by REML
Formula: radon ~ floor + (1 | county)
   AIC   BIC logLik deviance REMLdev
 2179 2199  -1086     2164    2171
Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 0.108    0.328
 Residual             0.571    0.756
Number of obs: 919, groups: county, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept)   1.4616     0.0516   28.34
floor        -0.6930     0.0704   -9.84

Correlation of Fixed Effects:
      (Intr)
floor -0.288
```

Note that the average intercept is 1.46, but the intercepts, across counties, have a standard deviation of $\sigma_\alpha = 0.33$.

Varying Intercepts, Floor Predictor

We can call for estimates of the county level coefficients:

```
> coef(M1)
```

```
$county
   (Intercept)      floor
1    1.1915015 -0.6929905
2    0.9276037 -0.6929905
...
83   1.5716904 -0.6929905
84   1.5906371 -0.6929905
85   1.3862299 -0.6929905
```

We can examine the fixed and random effects separately:

```
> fixef(M1)
```

```
(Intercept)       floor
      1.462      -0.693
```

Next, we examine the random effects, the amount by which the intercept in a given county varies around the central value of 1.46.

Varying Intercepts, Floor Predictor

```
> ranef(M1)

$county
   (Intercept)
1  -0.27009244
2  -0.53399029
...
85 -0.07536403
```

Uncertainties in the Estimates

Gelman & Hill have added a nice pair of functions for examining standard errors quickly.

```
> se.fixef(M1)

(Intercept)       floor
    0.05157     0.07043

> se.ranef(M1)

$county
             [,1]
 [1,] 0.24778450
 [2,] 0.09982720
 [3,] 0.26228596
...
[85,] 0.27967312
```

Summarizing and Displaying the Fitted Model

We can access the components of the estimates and standard errors using list notation in R. For example, to get a 95% confidence interval for the slope (which, in this model, does not vary by county),

```
> fixef(M1)["floor"] + c(-2,2) * se.fixef(M1)["floor"]
```

```
[1] -0.8339 -0.5521
```

In extracting elements of the coefficients from `coef()` or `ranef()`, we must first identify the grouping (`county` in this case). For example, here is the 95% CI for the intercept in county 26:

```
> coef(M1)$county[26,1] + c(-2,2)*se.ranef(M1)$county[26]
```

```
[1] 1.219 1.507
```

Varying Slopes, Fixed Intercept

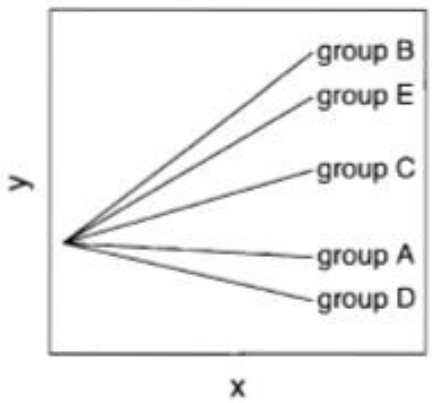
Another option is to let the slopes vary, while keeping a constant intercept This model may be written as

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \tag{8}$$

with

$$\beta_{j[i]} \sim \mathrm{N}(\mu_\beta, \sigma^2_\beta) \tag{9}$$

Here is a plot of this model:



Varying Slopes, Fixed Intercept

Fitting this model with `lmer()` is as follows:

```
> M2 ← lmer(radon ~ floor + (floor - 1 | county))
> display(M2)
```

```
lmer(formula = radon ~ floor + (floor - 1 | county))
            coef.est coef.se
(Intercept)  1.33     0.03
floor       -0.55     0.09

Error terms:
 Groups   Name  Std.Dev.
 county   floor 0.34
 Residual       0.81
---
number of obs: 919, groups: county, 85
AIC = 2258.8, DIC = 2234
deviance = 2242.5
```

Varying Slopes, Fixed Intercept

As before, we can examine individual coefficients:

```
> coef(M2)
```

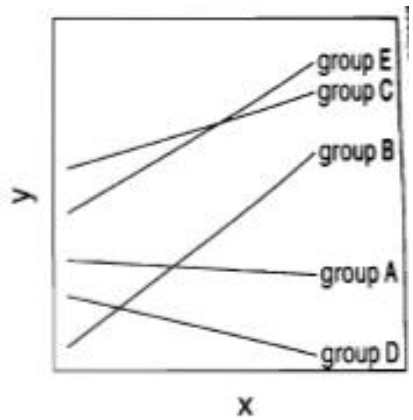
```
$county
   (Intercept)       floor
1     1.326744 -0.5522006
2     1.326744 -0.9269289
3     1.326744 -0.5361960
:        :          :
84    1.326744 -0.5455763
85    1.326744 -0.5546372
```

Varying Slopes, Varying Intercepts

Here is a model where the intercept *and* slope vary by group:

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \tag{10}$$

In this model, not only do the α and β coefficients have estimated standard errors, but they are also allowed to correlate across counties. (See p. 279 of Gelman & Hill.) Here is a plot of this model:



Varying Slopes, Varying Intercepts

Fitting this model goes like this:

```
> M3 ← lmer(radon ~ floor + (1 + floor | county) )
> display(M3)
```

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

Error terms:
 Groups   Name        Std.Dev. Corr
 county   (Intercept) 0.35
          floor       0.34     -0.34
 Residual             0.75
---
number of obs: 919, groups: county, 85
AIC = 2180.3, DIC = 2154
deviance = 2161.1
```

Varying Slopes, Varying Intercepts

Now, of course, we see differing slopes and intercepts across counties.

```
> coef(M3)

$county
   (Intercept)       floor
1    1.1445240 -0.5406161
2    0.9333816 -0.7708545
3    1.4716889 -0.6688832
:       :           :
84   1.5991210 -0.7327245
85   1.3787927 -0.6531793
```