Regression in Simple Multilevel Models

Nambari Shortcourse

15 July 2019

Overview

- Elaborate model for radon data to include predictors
  - Household-level predictors
  - County-level predictors
- Specify and fit models to the radon data
- Interpret regression coefficients
- Show how the model can be written in different ways

Radon data from GH Chapter 12

These data contain radon levels measured in one or more houses in each of 85 counties in Minnesota. See GH Section 1.2 for more details. We will just be looking at the radon levels in each county.

$$
\begin{aligned}
Y_{ij} &= \text{log radon level in house } i\text{, county } j \\
J &= \text{number of counties (85)} \\
j &= 1, \ldots, 85 \\
n_j &= \text{number of houses measured in county } j \\
i &= 1, \ldots, n_j
\end{aligned}
$$

This is a standard two-level structure

- Level 1 is the individual household
- Level 2 is the county
- Households are sampled within county

Two-level model assuming normal distribution

Level 1: Distribution of household radon level within county j

$$Y_{ij} \sim \mathcal{N}(\alpha_j, \sigma^2)$$

Level 2: Distribution of county-specific mean radon levels

$$\alpha_j \sim \mathcal{N}(\mu, \tau^2)$$

Incorporating a household-level predictor

In each household, radon was measured on the lowest floor.

$$\begin{aligned} X_{ij} &= 0 \text{ if basement} \\ &= 1 \text{ if first floor} \end{aligned}$$

In this notation, X_{ij} is the measurement location for household i in county j.

Pooled vs. multilevel regression

Pool data across counties

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta X_{ij}, \sigma^2)$$

In this model, there is one intercept and one slope.

County-specific intercepts

$$Y_{ij} \sim \mathcal{N}(\alpha_j + \beta X_{ij}, \sigma^2)$$

This model has a separate intercept for each county, but single slope for floor effect

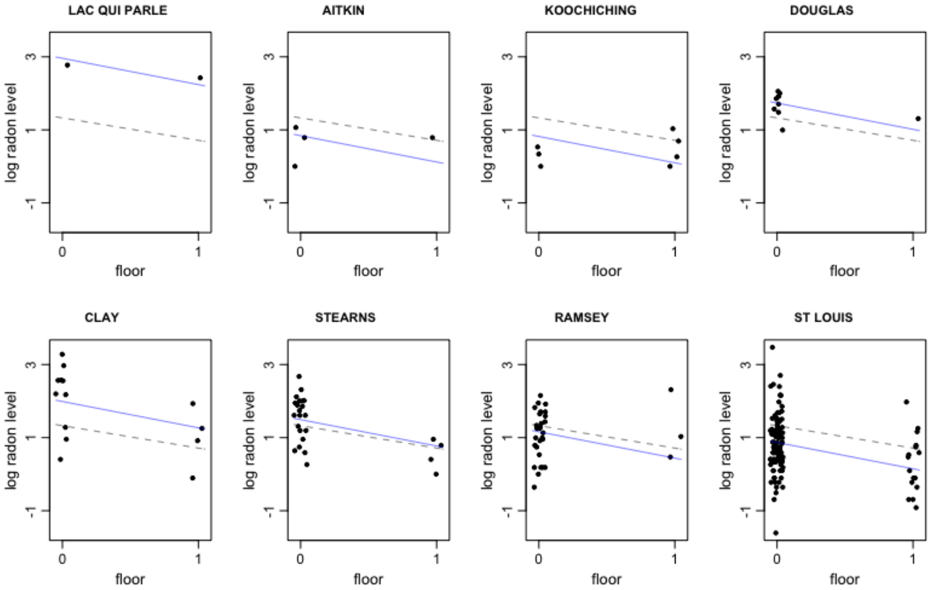
Pool data across counties

```
> ## Complete pooling regression
> lm.pooled <- lm (y ~ x)
> display (lm.pooled)
lm(formula = y ~ x)
            coef.est coef.se
(Intercept)  1.33     0.03
x           -0.61     0.07
---
n = 919, k = 2
residual sd = 0.82, R-Squared = 0.07
>
```

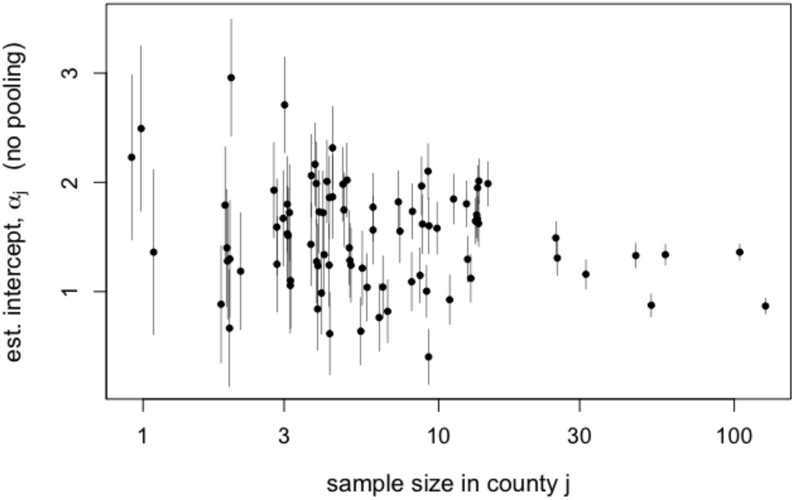
County-specific intercepts

```
> ## No pooling regression
> lm.unpooled <- lm (y ~ x + factor(county) -1)
> display (lm.unpooled)
lm(formula = y ~ x + factor(county) - 1)
                  coef.est coef.se
x                 -0.72     0.07
factor(county)1    0.84     0.38
factor(county)2    0.87     0.10
[...]
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
>
```

Pooled regression vs. county-specific intercept regression



County-specific intercepts



Multilevel regression

Multilevel model

Level 1: Within county variation

$$Y_{ij} \sim \mathcal{N}(\alpha_j + \beta X_{ij}, \sigma^2)$$

Level 2: Between county variation

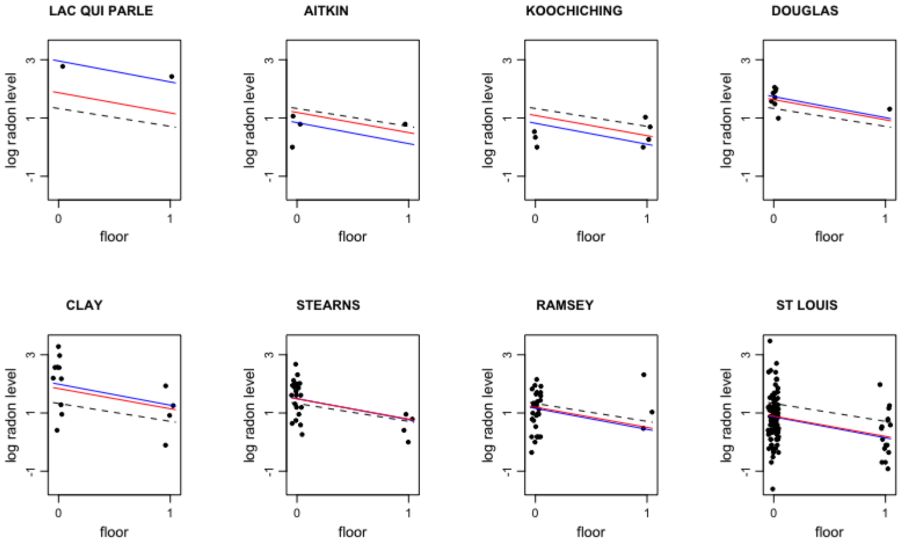
$$\alpha_j \sim \mathcal{N}(\mu, \tau^2)$$

Also has separate intercept for each county, but the variation is modeled explicitly.

Multilevel regression

```
> M1 <- lmer (y ~ x + (1 | county))
> display (M1)
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
---
number of obs: 919, groups: county, 85
AIC = 2179.3, DIC = 2156
deviance = 2163.7
>
```

Comparing county-level regressions for all 3 models



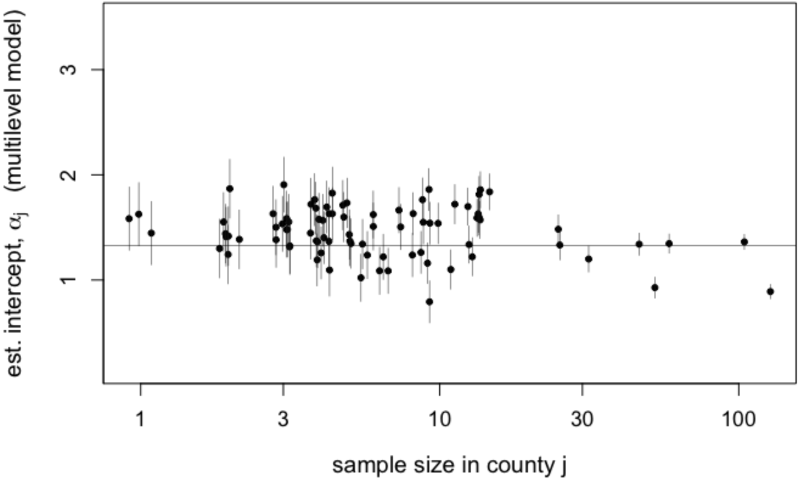
Estimate of intercepts from multilevel model

The estimates of α_j use both individual- and pooled regression estimates, and 'shrink' toward the pooled regression estimate depending on sample size and τ^2.

$$\widehat{\alpha}_j \approx \frac{(n_j/\sigma^2)(\overline{Y}_j - \beta\overline{X}_j) + (1/\tau^2)\mu}{(n_j/\sigma^2) + (1/\tau^2)}$$

What does this estimate correspond to when:
$\tau^2 \to 0$?
$\tau^2 \to \infty$?

Estimated intercepts from multilevel model



Including county-level predictor

- Multilevel models naturally incorporate predictors at the county level.
- It's less obvious how to do this when using indicators for the counties (model identifiability issues).
- In multilevel model, can just include county-level predictor into Level 2 of the model

Multilevel model with county-level predictor

Here we include U_j = county-level soil uranium level (log scale)

Level 1: Within county variation

$$Y_{ij} \sim \mathcal{N}(\alpha_j + \beta X_{ij}, \sigma^2)$$

Level 2: Between county variation

$$\alpha_j \sim \mathcal{N}(\mu + \gamma U_j, \tau^2)$$

Fitted model

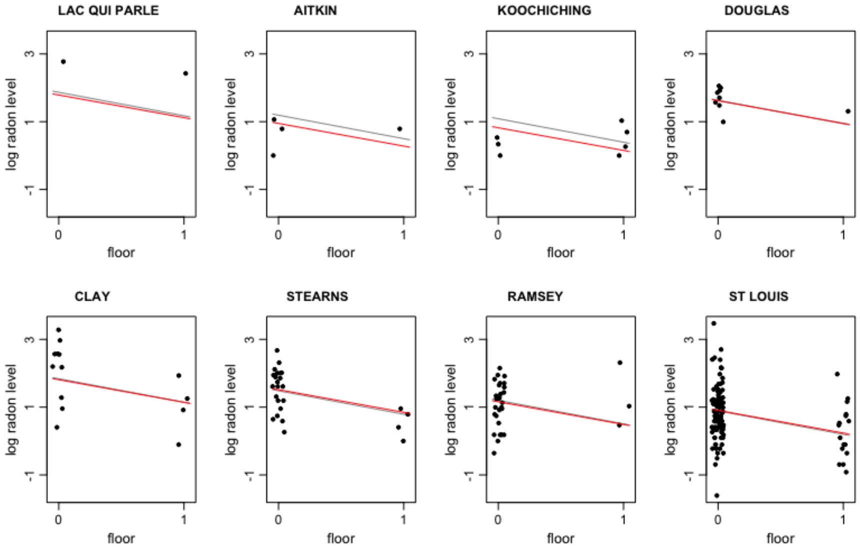
```
> M2 <- lmer (y ~ x + u.full + (1 | county))
> display (M2)
lmer(formula = y ~ x + u.full + (1 | county))
            coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.07
u.full       0.72     0.09

Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.16
 Residual             0.76
---
number of obs: 919, groups: county, 85
AIC = 2144.2, DIC = 2111.4
deviance = 2122.8
```

Estimated model coefficients

```
> coef (M2)
$county
   (Intercept)          x    u.full
1     1.445120 -0.6682448 0.7202676
2     1.477009 -0.6682448 0.7202676
3     1.478185 -0.6682448 0.7202676
4     1.576891 -0.6682448 0.7202676
5     1.473999 -0.6682448 0.7202676
6     1.439566 -0.6682448 0.7202676
7     1.593872 -0.6682448 0.7202676
[...]
82    1.490002 -0.6682448 0.7202676
83    1.398639 -0.6682448 0.7202676
84    1.551352 -0.6682448 0.7202676
85    1.423816 -0.6682448 0.7202676
```

Fitted regressions from multilevel model

Red = model with uranium as predictor



Representation of uranium effect