ANOVA, Pooling and Simple Multilevel Models

Nambari Short Course

15 July 2019

Overview

- Describe and summarize radon data from Chapter 12 in GH
- Describe multiple sources of variation
- Specify and fit multilevel model for radon data
- Examine county-specific estimates from multilevel model

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

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

Distribution of Y_{ij}: Partitioning variance

We can consider the *conditional* and *marginal* means and variances

- *Conditional* mean/variance is within county
- *Marginal* mean/variance is across counties

Helps focus on specific quantities in the model, and how to draw inference about them.

Conditional distribution

Within county, Y_{ij} has mean α_j and variance σ^2:

$$\begin{aligned} E(Y_{ij} \mid \alpha_j) &= \alpha_j \\ \text{var}(Y_{ij} \mid \alpha_j) &= \sigma^2 \end{aligned}$$

The *conditional variance* captures the variability of individual-specific household measures around the county-specific mean.

Marginal distribution

To determine marginal means and variances, we need to average across county.

Marginal mean

$$\begin{aligned} E(Y_{ij}) &= E\left\{E(Y_{ij} \,|\, \alpha_j)\right\} \\ &= E\left\{\alpha_j\right\} \\ &= \mu \end{aligned}$$

The overall mean of the Y_{ij} is the overall mean aggregated over counties.

Marginal distribution

Marginal variance

$$\begin{aligned} \mathrm{var}(Y_{ij}) &= E\left\{\mathrm{var}(Y_{ij} \mid \alpha_j)\right\} + \mathrm{var}\left\{E(Y_{ij} \mid \alpha_j)\right\} \\ &= E\left\{\sigma^2\right\} + \mathrm{var}\left\{\alpha_j\right\} \\ &= \sigma^2 + \tau^2 \end{aligned}$$

The overall variance of an individual radon measurement has two sources:

- variation of measures within specific county (σ^2)
- variation of county-specific means around the overall mean (τ^2)

Marginal correlation and covariance

Within each county, we assume the household specific measures are independent of each other.

However, when looking at the entire sample as a whole, measurements within county are correlated.

$$
\begin{aligned}
\text{cov}(Y_{ij}, Y_{i'j}) &= \tau^2 \\
\text{corr}(Y_{ij}, Y_{i'j}) &= \frac{\tau^2}{\tau^2 + \sigma^2}
\end{aligned}
$$

What does this mean when:

- Within-county variation is large relative to between-county variation?
- Between-county variation is large relative to within-county variation?

Estimating parameters from a multilevel model

Possible objectives

- Overall mean, standard error
- County-specific estimates, standard errors
    - Gives a sense of variation across counties

Estimating parameters from a multilevel model

Overall mean

Could just take the mean of the pooled data

$$\overline{Y}_{\text{all}} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j=1}^{J} n_j}$$

This is a valid estimate but it's kind of over-simplistic because it ignores variation between counties.

Estimating parameters from a multilevel model

County-specific means
Could use sample mean for each county

$$\overline{Y}_j = (1/n_j) \sum_{i=1}^{n_j} Y_{ij}$$

Can also calculate county-specific standard errors in the usual way

- What are the pluses and minuses of this approach
- Think about the varying sample sizes within each county

Estimating parameters from a multilevel model

Pooled estimates of county-specific means In multilevel models, we can use information about both the overall mean and the county-specific means to construct a *partially pooled* estimate

$$\hat{\alpha}_j \approx \frac{(n_j/\sigma^2)\overline{Y}_j + (1/\tau^2)\overline{Y}_{\text{all}}}{(n_j/\sigma^2) + (1/\tau^2)}$$

Notice that this is a weighted average of the county-specific sample means and the overall sample mean.

- What happens for counties where n_j is small? large?
- How do within- and between-county variation affect how the weighting is done?

Analysis of Radon Data

Model 1

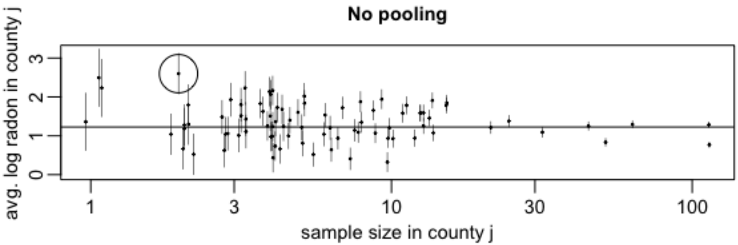
$$Y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$$

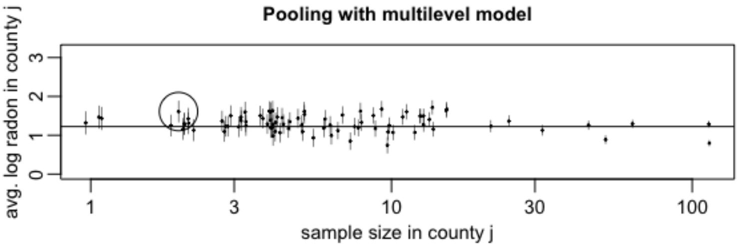
Model 2

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

Model 3

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}(\alpha_j, \sigma^2) \\ \alpha_j &\sim \mathcal{N}(\mu, \tau^2) \end{aligned}$$


No pooling
avg. log radon in county j
0
1
2
3
1
3
10
30
100
sample size in county j


Pooling with multilevel model
avg. log radon in county j
0
1
2
3
1
3
10
30
100
sample size in county j