SOC6707 Intermediate Data Analysis

Monica Alexander

Week 9: Introduction to hierarchical models

Hierarchical models

- Hierarchical models used to estimate parameters in settings where there is a hierarchy of nested populations.
- Many problems have a natural hierarchy e.g.
  - patients within hospitals
  - school kids within classes within schools
  - maternal deaths within countries within regions within the world
- Want to get estimates of underlying parameters of interest (e.g. probability of dying, test score, risk of disease) accounting for the hierarchy in the data
- A natural framework for including information at different levels of the hierarchy

Radon example

- Radon is a naturally occurring radioactive gas.
- Its decay products are also radioactive; in high concentrations, they can cause lung cancer (several 1000 deaths/year in the USA).
- Radon levels vary greatly across US homes.
- Data: radon measurements in over 80K houses throughout the US.
- Hierarchy: houses observed in counties.
- Potential predictors: floor (basement or 1st floor) in the house, soil uranium level at country level.

Radon dataset

Selected rows and columns

idnum	state	county	basement	activity
1	AZ	APACHE	N	0.3
2	AZ	APACHE		0.6
3	AZ	APACHE	N	0.5
4	AZ	APACHE	N	0.6
5	AZ	APACHE	N	0.3
6	AZ	APACHE	N	1.2

- 12,777 observations from 386 counties

What might we want to estimate/predict?

What might we want to estimate/predict?

- Expected radon level in a county
- Expected radon level in a county we did not have samples for
- Predicted radon level for a newly observed house in a particular county
- . . . ?

Let's introduce some notation

- units $i = 1, \ldots n$, the smallest items of measurement (household)
- outcome $y = (y_1, \ldots, y_n)$. The unit-level outcome being measure (log radon)
- groups $j = 1, \ldots, J$ (counties)
- Indexing $j[i]$ (the county for house i)
- x_i is an indicator, whether or not measurement was taken on basement (house level)
- u_j is the uranium level in the soil (county level)

Notation

Thinking about our usual regression set-up, we usual write as something like

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Let's rewrite this as

$$Y_i \sim N(\mu_i, \sigma^2)$$

with

$$\mu_i = \beta_0 + \beta_1 X_i$$

and

$$\varepsilon_i \sim N(0, \sigma^2)$$

These are equivalent.

A model for log radon

$$Y_i \sim N(\mu_i, \sigma^2)$$

- Note that $\mu_i = E(Y_i)$ i.e. the expected (log) radon level for a particular house i
- How to model μ_i?
- Let's start simple (no covariates)
- Given we know house i is in county j, how can we model μ_i?

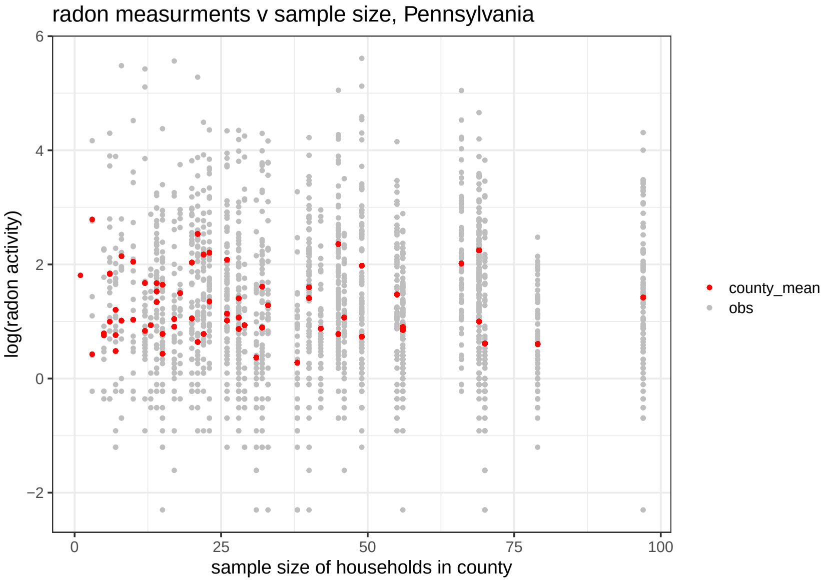
One option: no pooling

Estimate the county-level mean for each county, using only the data from that county. The model is

$$y_i \sim N\left(\alpha_{j[i]}^{\text{no pool}}, \sigma_y^2\right)$$

- The "no pool" refers to treating each county separately, i.e. no pooling of information across counties
- The most appropriate estimator for this is the county mean, i.e. $\bar{y}_j$
- I.e. the expected level of log radon for a particular house i in county j is just the mean radon level for the county

No pooling



What do you notice about this graph?

Another option: complete pooling

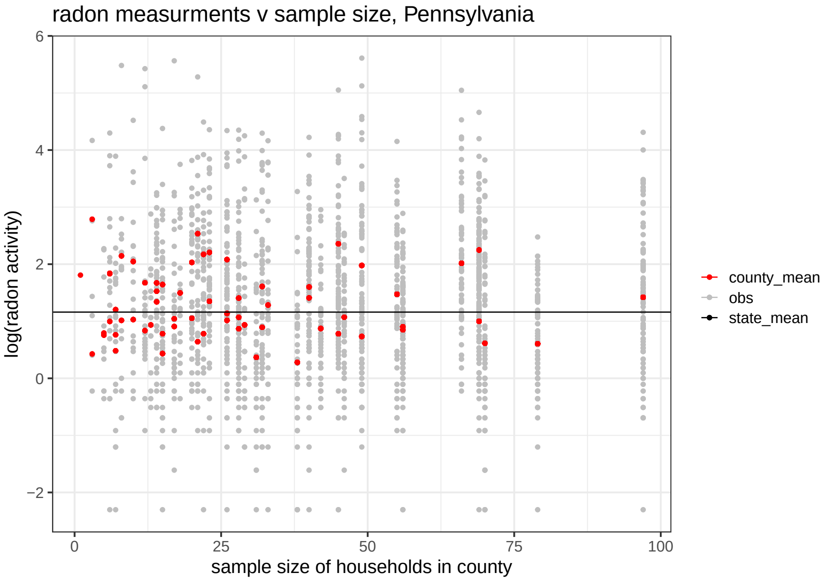
- Maybe we believe that the expected radon level for a particular house is not going to vary by county
- Use the state mean as the best estimate for the means in each county.

Model is

$$y_i \sim N\left(\mu, \sigma_y^2\right)$$

- i.e. expected log radon level is constant across state
- Best estimator here would just be state mean
- this is referred to "complete pooling" because information across all counties is pooled together

Complete pooling



Pros? Cons?

A happy medium

- Ideally we want to allow expected county radon levels to differ
- But we also want to account for information across all counties and not treat counties as separate
- A solution: partial pooling via hierarchical modeling

Another option: hierarchical model

- The expected radon level in a particular house *i* is
- county means α_j come from some common distribution across a state
- there are some underlying parameters governing the distribution of α's, which are generally unknown
- middle ground between first two options, α's are similar but not the same

Hierarchical model

The model is

$$\begin{aligned} y_i &\sim N\left(\alpha_{j[i]}, \sigma_y^2\right) \\ \alpha_j &\sim N\left(\mu_\alpha, \sigma_\alpha^2\right) \end{aligned}$$

- The $alpha_j$'s are themselves assumed to be from a common distribution
- μ_α and σ_α are called **hyperparameters**

Hierarchical model

$$\begin{tikzcd}
 & \mu_\alpha, \sigma_\alpha \arrow[dl] \arrow[d] \arrow[drr] & & \\
\alpha_1 \arrow[d] & \alpha_2 \arrow[d] & \cdots & \alpha_n \arrow[d] \\
y_1, y_2 & y_3, \ldots y_8 & & y_{12770}, \ldots y_{12777}
\end{tikzcd}$$

Because of the hierarchical set-up, the resulting estimates for the county means are in-between the no-pooling and complete-pooling estimates.

Compare to

- No pooling

$$\begin{tikzcd}[column sep=small]
\alpha_1 \arrow[d] & \alpha_2 \arrow[d] & \cdots & \alpha_n \arrow[d] \\
y_1, y_2 & y_3, \dots y_8 & & y_{12770}, \dots y_{12777}
\end{tikzcd}$$

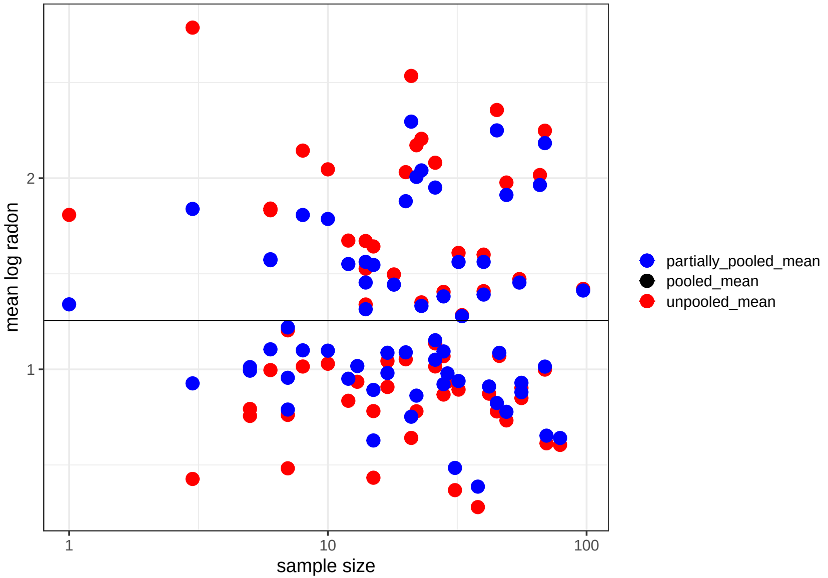
- Complete pooling

$$\begin{tikzcd}
& \mu \arrow[dl] \arrow[d] \arrow[drr] & & \\
y_1 & y_2 & \cdots & y_{12777}
\end{tikzcd}$$

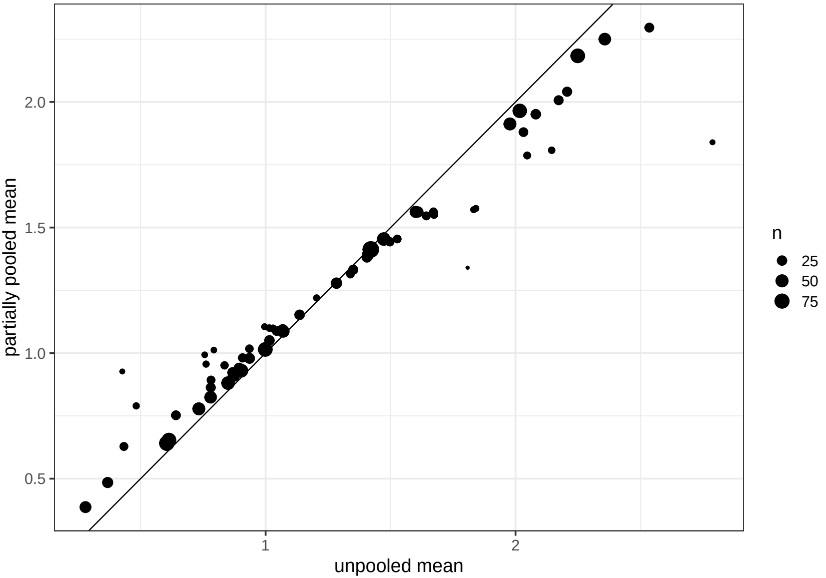
Many names

- Also known as multilevel models, I will probably flip between the two
- Fixed and random effects
  - α_j's commonly referred to as random effects, because they are modeled as random variables
  - fixed effects are parameters that don't vary by group, or to parameters that vary but are not modeled themselves (e.g. county/state indicator variables)
- random effects models, (generalized) linear mixed models, mixed effects models: often used as synonyms for multilevel models

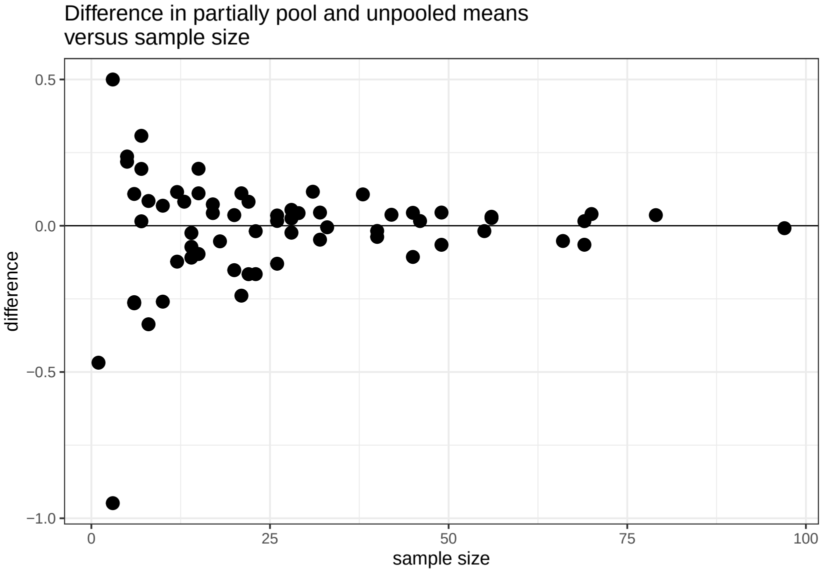
The effect of partial pooling in the radon case



The effect of partial pooling in the radon case



The effect of partial pooling in the radon case



Where are we at

- Hierarchical models allow for ‘information exchange’ across groups
- Has the effect ‘shrinking’ group means to the overall mean
- Shrinking effect is larger when the sample size in a particular group is smaller

Why does this happen?

- It turns out that the estimate of the hierarchical mean $\hat{\alpha}_j$ is a weighted mean between information from that group j and all the other groups:

$$\hat{\alpha}_j = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}$$

Adding covariates

Adding covariates

For the radon example:

- The measurements are not exactly comparable across houses because in some houses, measurements are taken in the basement, while in other houses, 1st floor measurement are taken. (This is x_i)
- Additionally, county-level uranium measurements are probably informative for across-county differences in mean levels. (This is u_j)

When adding covariates, need to think about

- what level the covariate relates to
- whether or not to model the effect hierarchically

Including covariates at the unit level

- Let x_i be the house-level first-floor indicator (with $x_i = 0$ for basements, 1 otherwise).
- This is a house-level covariate
- We can include house-level predictors in the house-level mean as follows:

$$
\begin{aligned}
y_i &\sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n \\
\alpha_j &\sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J
\end{aligned}
$$

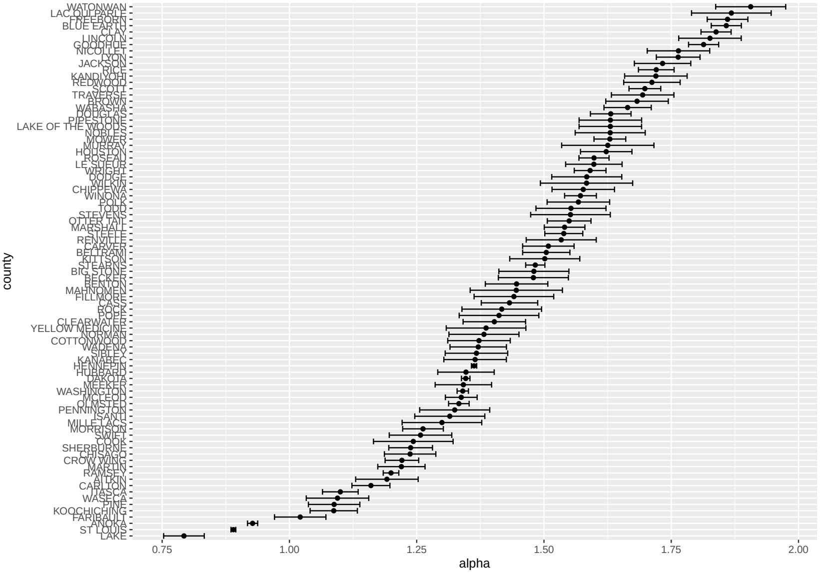
Note: we have varying intercepts but a constant slope

Covariates at unit level

$$y_i \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n$$
$$\alpha_j \sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J$$

- Estimate of β is -0.693
- Estimate of μ_α is 1.462

County-specific intercepts



Including covariates at the group level

- County-level log-uranium measurements u_j are probably informative for across-county differences in mean levels.
- We can include group-level predictors in the group-level mean as follows:

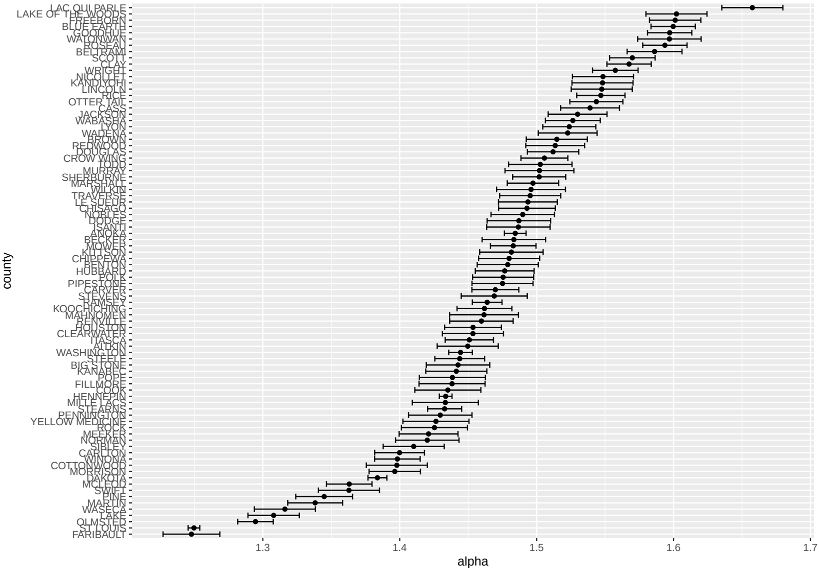
$$y_i \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n$$
$$\alpha_j \sim N\left(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J$$

Adding covariates at group level

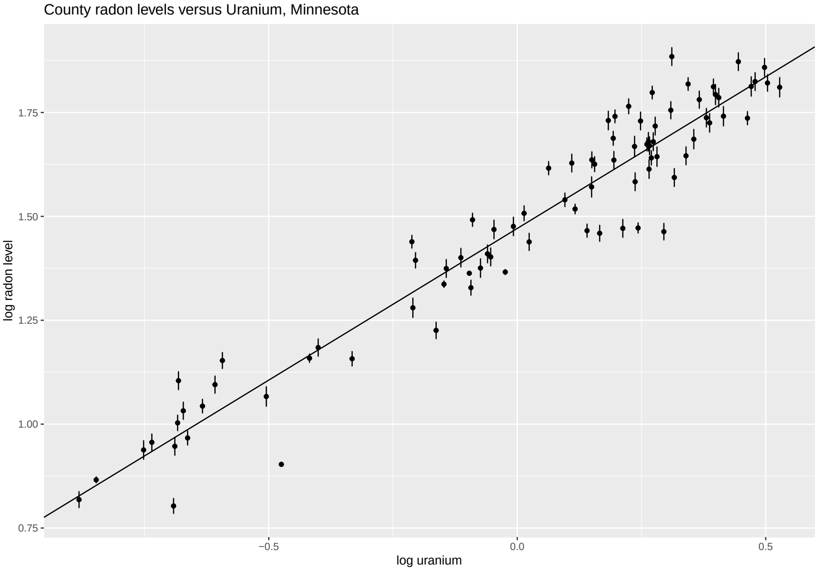
$$y_i \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n$$
$$\alpha_j \sim N\left(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J$$

- Estimate of β is -0.668
- Estimate of γ_0 is 1.407
- Estimate of γ_1 is 0.729

County-specific intercepts



County-level radon and uranium



Varying slopes (TBD)

What about letting the effect of x_i vary by county?

- In last model, we assume that the difference between basement and first floor measurement is the same across houses, no matter which county the house is in.
- What if that difference varies by county?

$$
\begin{aligned}
y_i &\sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n \\
\alpha_j &\sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J \\
\beta_j &\sim N\left(\mu_\beta, \sigma_\beta^2\right), \text{ for } j = 1, 2, \ldots, J
\end{aligned}
$$

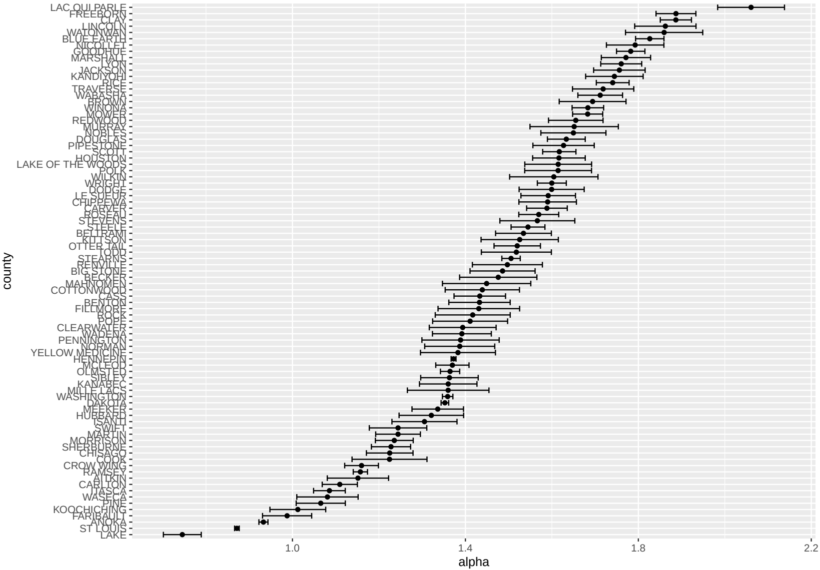
Allowing for varying slopes.

Allowing for varying slopes at unit level

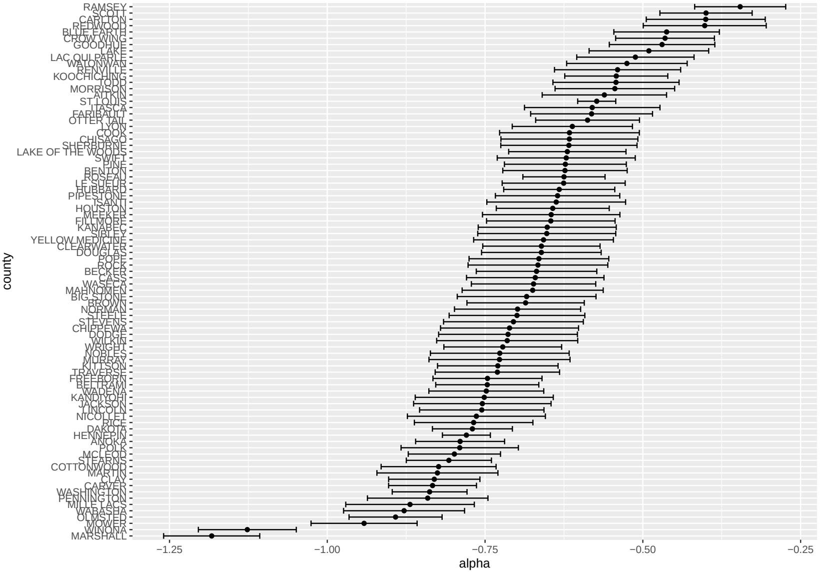
$$y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right), \text{ for } i = 1, 2, \ldots, n$$
$$\alpha_j \sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \text{ for } j = 1, 2, \ldots, J$$
$$\beta_j \sim N\left(\mu_\beta, \sigma_\beta^2\right), \text{ for } j = 1, 2, \ldots, J$$

- Estimate of μ_α is 1.32
- Estimate of μ_β is -0.539

County-specific intercepts



County-specific slopes



Hierarchical models in R

Fitting hierarchical models in R

- Many different options and packages to do this
- Many powerful options fitting Bayesian hierarchical models using languages like Stan or JAGS (but no time!)
- We will be using the `lme4` package, which allows you to fit hierarchical models using commands that are a logical extension of `lm` and `glm`
- (So you will need to `install.packages(lme4)`)

Radon levels in Minnesota

We will see in lab, but a brief introduction to notation.

What we would usually do:

```r
library(lme4)
d_mn <- d %>% filter(state=="MN")

mod_nopool <- lm(log_activity ~ county, data = d_mn)
mod_pool <- lm(log_activity ~ 1, data = d_mn)
```

Hierarchical model:

```r
mod_hier <- lmer(log_activity ~ (1 | county), data = d_mn)
```

Radon levels in Pennsylvania

Adding covariates:

```
mod_hier <- lmer(log_activity ~ floor + log_uran + (1 | county), data = d_mn)
```