STA2201H Methods of Applied Statistics II

Monica Alexander

Week 8: Hierarchical models

Notes

- A2 due soon

Hierarchical models

Reading

- BDA Chapter 5
- GH Chapters 11-15

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
- Hierarchical models are a natural framework for including information at different levels of the hierarchy
- Particularly useful when there is little information about some groups

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.

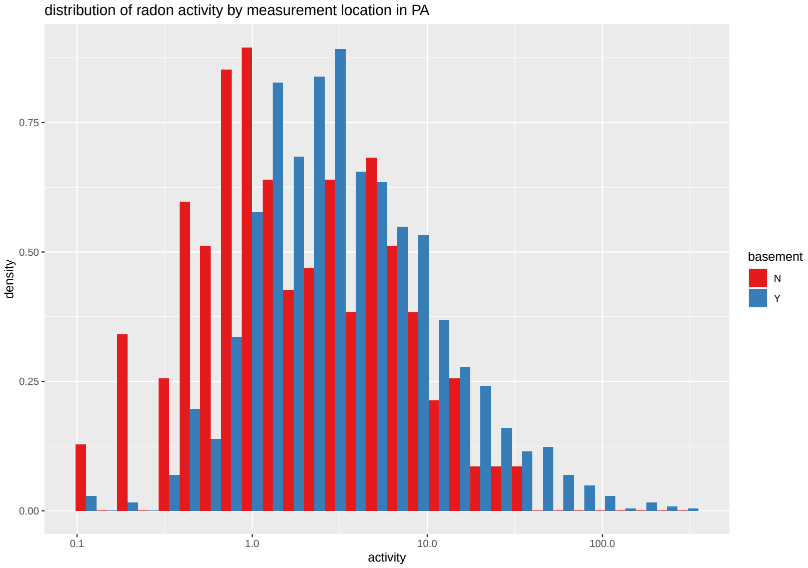
Radon dataset

Selected rows and columns

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

- 12,777 observations from 386 counties

Radon dataset



Research questions

- What's the expected level of radon in a particular county for which we observe data?
- What's the predicted level of radon in a particular county for which we don't have data?
- What's the predicted level of radon for a not-yet-sampled house?
- What's the effect on radon level if we take the measurement on the first floor compared to in the basement?

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)
- ... we may need second level of groups $k = 1, \ldots, K$ e.g. states
- Indexing $j[i]$ (the county for house i)
- $\bar{y}_j = 1/n_j \sum_{i \in G_j} y_i$ is the group mean (county mean)

Radon likelihood

Let's assume the y_i's are normally distributed and conditionally independent

$$y_i|\mu_i, \sigma_y^2 \overset{i.i.d}{\sim} N\left(\mu_i, \sigma_y^2\right)$$

- how to model groups means?
- what expression to use for μ_i?

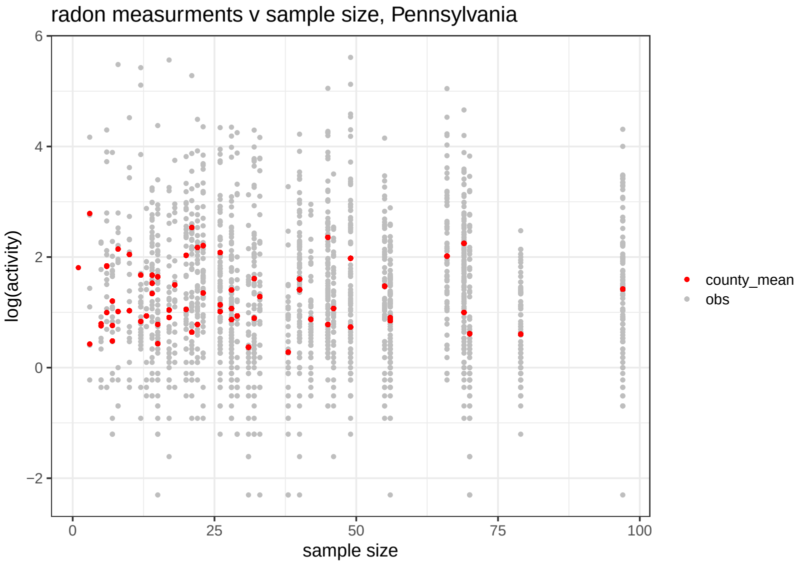
One option: no pooling

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

$$y_i|\alpha_{j[i]}^{nopool}, \sigma_y^2 \sim N\left(\alpha_{j[i]}^{nopool}, \sigma_y^2\right)$$

- the MLE would just been the sample means i.e. $\bar{y}_j$

No pooling



Pros? Cons?

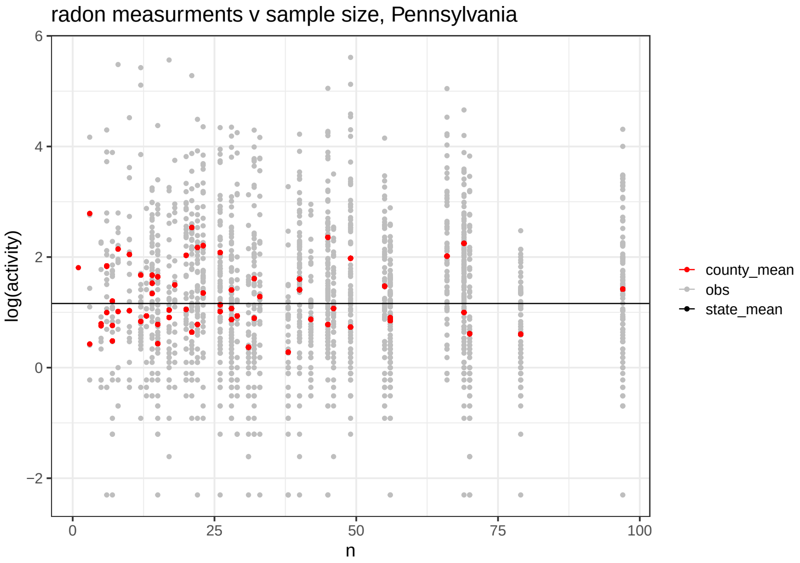
Another option: complete pooling

Use the state mean as the best estimate for the means in each county.

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

Again, frequentist estimate would just be state mean

Complete pooling



Pros? Cons?

Another option: hirerachical model

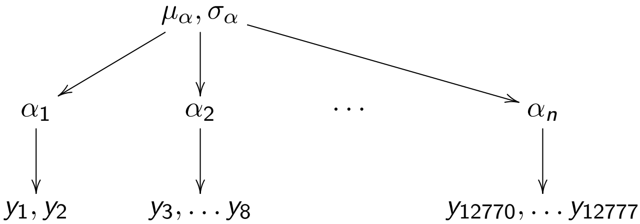
- county means α_j come from some common distribution across a state
- there are some underlying parameters governing the distribution of αs, which are generally unknown
- middle ground between first two options, αs are similar but not the same
- c.f. bias variance trade-off

Write model as

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

μ_α and σ_α are called **hyperparameters**. We've seen these before! This is looking very Bayesian!

Hierarchical model



$\mu_\alpha, \sigma_\alpha$

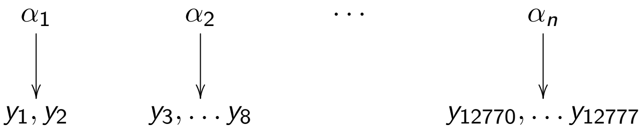
α_1 α_2 $\cdots$ α_n

y_1, y_2 $y_3, \ldots y_8$ $y_{12770}, \ldots y_{12777}$

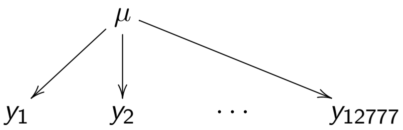
Because of the hierarchical set-up, the resulting estimates for the county means are in-between the no-pooling and complete-pooling estimates.

Compare to

- No pooling



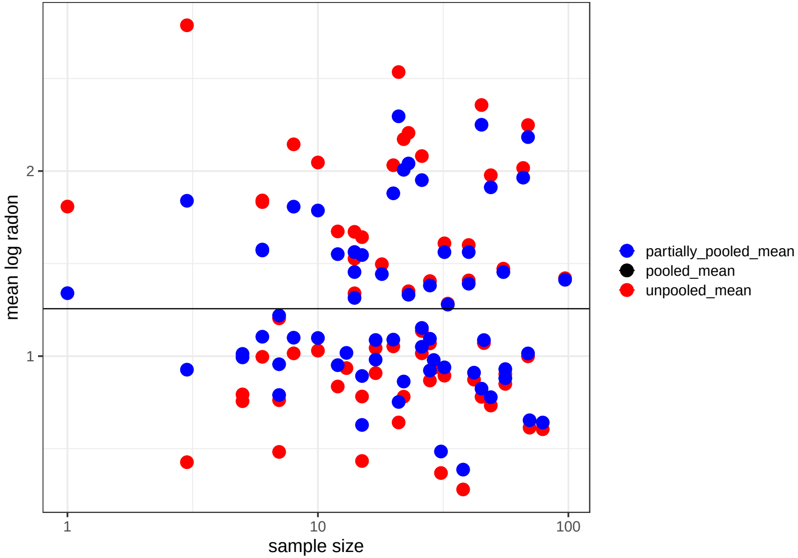
- Complete pooling



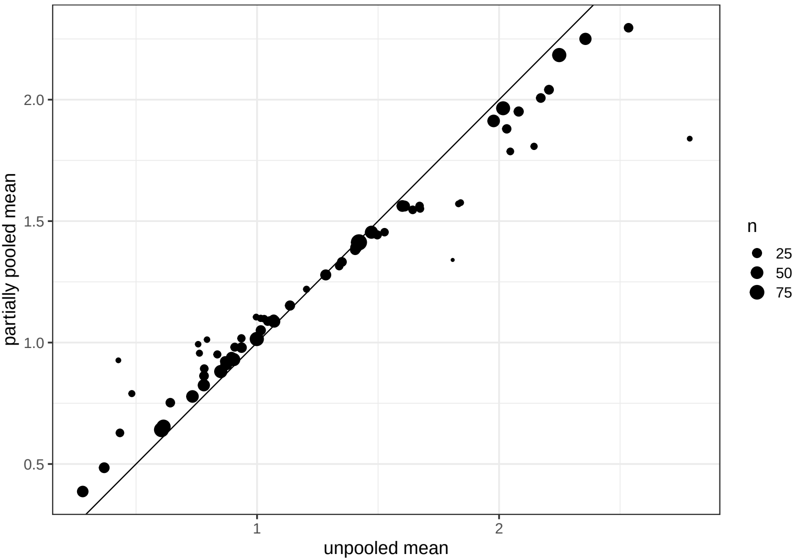
Hierarchical models? Or something else, depending on what field you're in

- 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

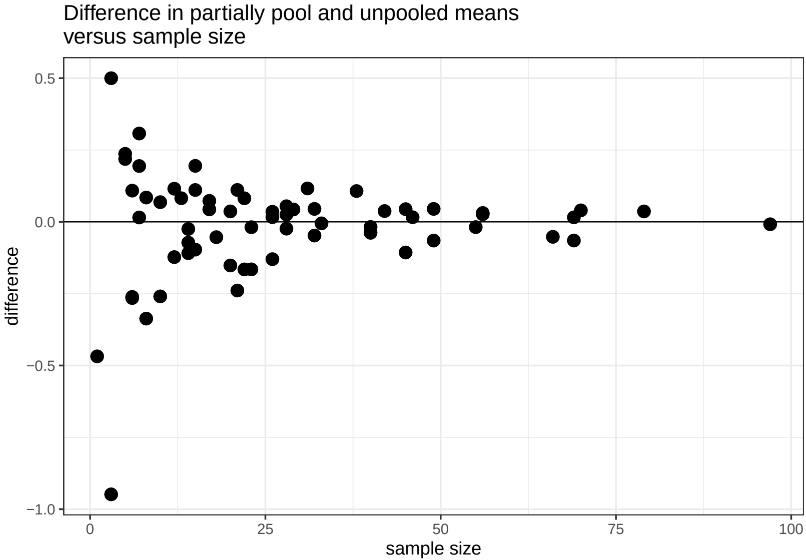
The effect of partial pooling in the radon case



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

What does a partially pooled mean look like?

For the model

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

The conditional distribution for α_j is

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

How was this obtained? Bayes rule.
$p\left(\alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha\right) \propto p\left(\boldsymbol{y}|\alpha_j, \sigma_y\right) p\left(\alpha_j|\mu_\alpha, \sigma_\alpha\right)$

We've seen this story before. $\hat{\alpha}_j$ is a weighted mean.

Hierarchical models in a Bayesian context

More general notation

- Interested in outcome y
- y depends on parameters θ
- θ itself depends on parameters ϕ

The key 'hierarchical' part of these models is that ϕ is not known and thus has its own prior distribution, $p(\phi)$.

In the radon set up, θ is $[\alpha_j, \sigma_y]$ and ϕ is $[\mu_\alpha, \sigma_\alpha]$

Going full Bayes

- We are incorporating uncertainty about ϕ in the model through specifying a prior distribution
- The joint prior distribution is $p(\phi, \theta) = p(\theta|\phi)p(\phi)$
- The joint posterior distribution is $p(\phi, \theta|y) \propto p(\phi, \theta)p(y|\phi, \theta)$

Priors on hyper-parameters

The same old story as in non-hierarchical models:

- "it is often practical to start with a simple, relatively non-informative, prior distribution on ϕ and seek to add more prior information if there remains too much variation in the posterior distribution." (BDA pg 108)
- Recommendations change
- Stan group recommendations here: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- Related to priors on scale parameters, recommendation is half-normal(0,1) or half-t(4,0,1). Earlier recommendations (e.g. in GH) may be too spread out, placing too much mass on cases with minimal pooling
- when in doubt: check sensitivity, plot plot plot

A short aside: exchangeability

Exchangeability

- y_i may be regarded as exchangeable if we express uncertainty as a joint probability density $p(y_1, \ldots, y_n)$ that is invariant to permutations of the indexes.
- When levels are exchangeable, we can reassign the indices arbitrarily and lose no information; that is, the joint distribution will remain the same
- ‘Ignorance implies exchangeability’

Exchangeability: hierarchical models

Our goal is to estimate some parameter(s) θ, which we are modeling hierarchically such that θ is governed by some ϕ.

- In the hierarchical set-up, if no information, other than the data y, are available to distinguish any of the θ_js, the $(\theta_1, \ldots, \theta_J)$ are exchangeable and their joint distribution is invariant to permutations of the indexes.
- The θ's are assumed to be 'similar' in the sense that the labels convey no information
- This is equivalent to to assuming that θ_js are drawn from a common prior distribution with some unknown parameters

When you have additional information

Often observations are not fully exchangeable, but are *partially* or *conditionally* exchangeable:

Partial exchangeability:

- If the observations can be grouped (a priori) then use hierarchical model
- In hierarchical models, we treat the levels of the group as exchangeable, and observations within each level in the group as also exchangeable.
- e.g. houses within counties are exchangeable, and counties are exchangeable

When you have additional information

Conditional exchangeability:

- If y_i has additional information x_i so that y_i are not exchangeable but (y_i, x_i) still are exchangeable, then we can make a joint model for (y_i, x_i) or a conditional model for $y_i|x_i$
- So in the radon case, measurements within a county y_i may not be exchangeable, but pairs of radon / basement observations (y_i, x_i) may be exchangeable

Exchangeability

Exchangeability justifies why we can use

- a joint model for data
- a joint prior for set of parameters

An assumption of exchangeability is a judgement based on our knowledge of the context.

Back to radon

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

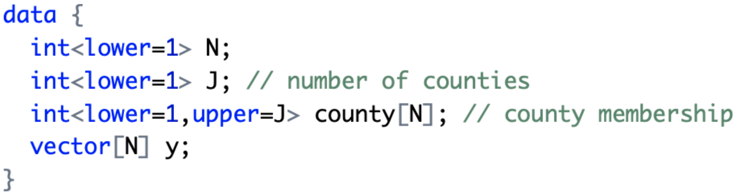
Let's put some priors on the hyperparameters and on σ_y:

$$\begin{gathered} \sigma_y \sim N^+(0,1) \\ \sigma_\alpha \sim N^+(0,1) \\ \mu_\alpha \sim N(0,1) \end{gathered}$$

How to run in Stan?

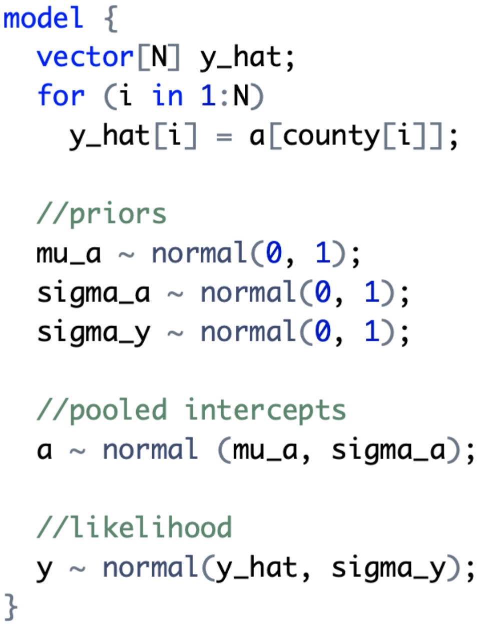
- We need to input additional information about group membership
- As well as the usual y, N, X inputs, we need things like
  - J: number of groups (counties)
  - "*group.i*": the group membership of observation i (e.g. which county household i is in).
    - what is the length of this?
    - note that this must be an integer (e.g. can't just put in county names)

Stan indexing



```
data {
  int<lower=1> N;
  int<lower=1> J; // number of counties
  int<lower=1,upper=J> county[N]; // county membership
  vector[N] y;
}
```

Stan indexing



```
model {
  vector[N] y_hat;
  for (i in 1:N)
    y_hat[i] = a[county[i]];

  //priors
  mu_a ~ normal(0, 1);
  sigma_a ~ normal(0, 1);
  sigma_y ~ normal(0, 1);

  //pooled intercepts
  a ~ normal (mu_a, sigma_a);

  //likelihood
  y ~ normal(y_hat, sigma_y);
}
```

Run in Stan

```
round(summary(mod1)$summary[1:10,c("mean", "se_mean", "n_eff", "Rhat")]
```

```
##       mean se_mean   n_eff Rhat
## a[1]  1.06    0.01  424.99 1.00
## a[2]  0.89    0.00  517.96 1.00
## a[3]  1.19    0.01  497.01 1.00
## a[4]  1.22    0.01  635.15 1.00
## a[5]  1.28    0.01  610.89 1.01
## a[6]  1.38    0.01  405.65 1.00
## a[7]  1.72    0.01  907.56 1.00
## a[8]  1.44    0.01  482.74 1.00
## a[9]  1.08    0.01 1040.82 0.99
## a[10] 1.25    0.01  555.91 0.99
```

Predicting new observations

- Question of interest: how to predict $\tilde{y}_k$ for a non-yet-sampled unit k in group $j[k]$, on which we may or may not have data?
- In this example, how do we predict the log radon level for a new house?

You should know the answer to this from last lecture!

Predicting new observations

Use the posterior predictive distribution

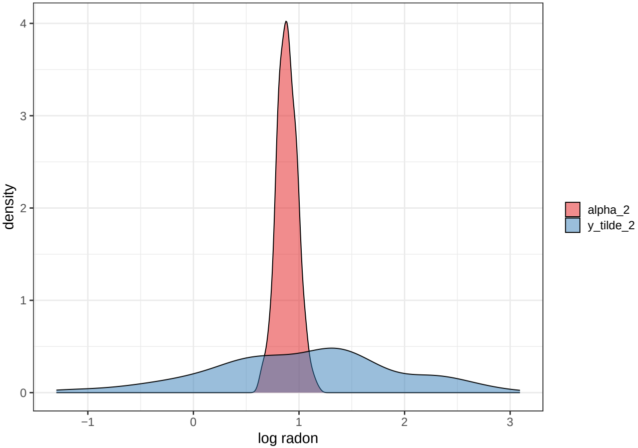
$$p\left(\tilde{y}_k|\boldsymbol{y}\right) = \int_{\boldsymbol{\theta}} p\left(\tilde{y}_k|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

In the radon case, $\theta = (\alpha_{j[k]}, \sigma_y^2)$

- Often hard to sample from $p\left(\tilde{y}_k|\boldsymbol{y}\right)$, so what do we do in practice?
  - Sample $\boldsymbol{\theta}^{(s)} \sim p(\boldsymbol{\theta}|\boldsymbol{y})$
  - Sample $\tilde{y}_k^{(s)} \sim p\left(\tilde{y}_k|\boldsymbol{\theta}^{(s)}\right)$,
  - In the radon case, $\tilde{y}_k|\alpha_{j[k]}^{(s)}, \left(\sigma_y^2\right)^{(s)} \sim N\left(\alpha_{j[k]}^{(s)}, \left(\sigma_y^2\right)^{(s)}\right)$

Predicting new observations

e.g. samples of a new observation from county 2 below, compared to the mean of county 2 for Minnesota. What's the difference between $p\left(\tilde{y}_k|\boldsymbol{y}\right)$ and $p\left(\alpha_{j[k]}|\boldsymbol{y}\right)$?



What if the county is not in the data set?

- Can we still get a prediction for a household in this county? Yes.
- E.g we do not have any observations for Red Lake county in Minnesota, call this county number 86
- What do we do?

What if the county is not in the data set?

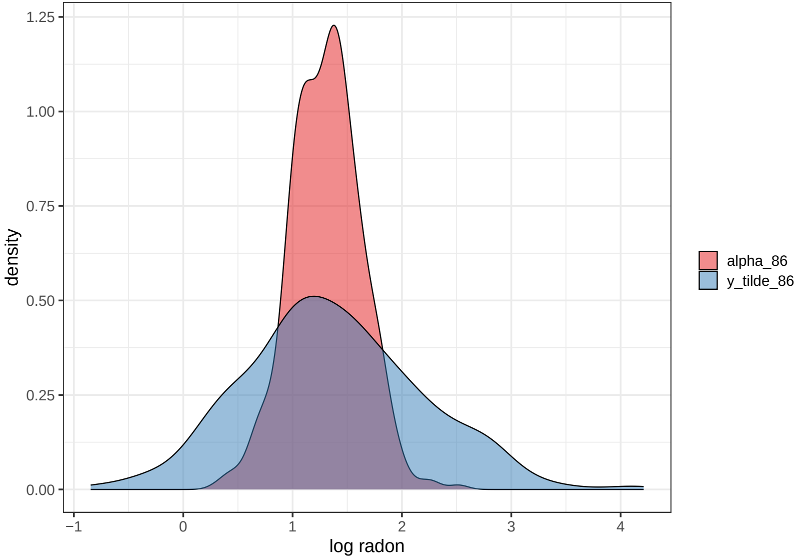
- We first need to sample and α_{86} from its (predictive) posterior distribution

$$p\left(\tilde{\alpha}_{j[k]}|\boldsymbol{y}\right) = \int_{\boldsymbol{\theta}} p\left(\tilde{\alpha}_{j[k]}|\boldsymbol{\theta}, \boldsymbol{y}\right) p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) d\boldsymbol{\theta}$$

... then proceed as before

- As previously, can do in Stan in generated quantities block, or post-fitting in R, using the posterior samples.

Observation from new county



Predicting new household from new county

- The ability to predict a new household from an unobserved county is an extremely useful feature of the hierarchical set-up
- c.f. maternal mortality project, we need estimates for 193 countries, we only have observations from around 150.
- But (as always) be careful of the assumptions you're making, and whether they're reasonable.

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

- Questions:
  - what are we inherently assuming about county 86?
  - based on answer to above, why do we have to simulate a new α_{86}?

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.
- Additionally, county-level uranium measurements are probably informative for across-county differences in mean levels.

Straight forward to add covariates to existing model, but need to think about

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

Including covaiates 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:

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

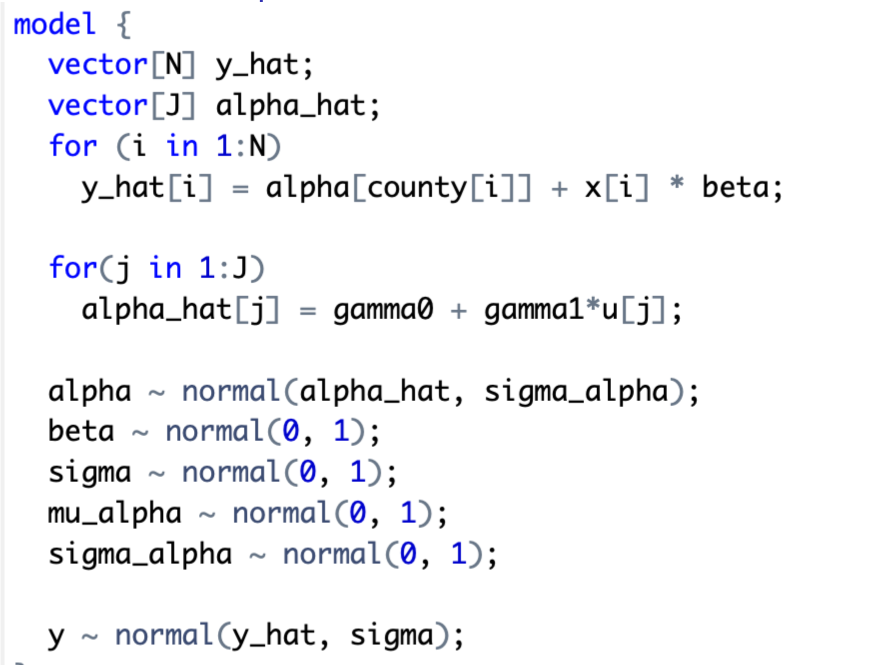
Note: we have varying intercepts but a constant slope

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:

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

Run in Stan - one option



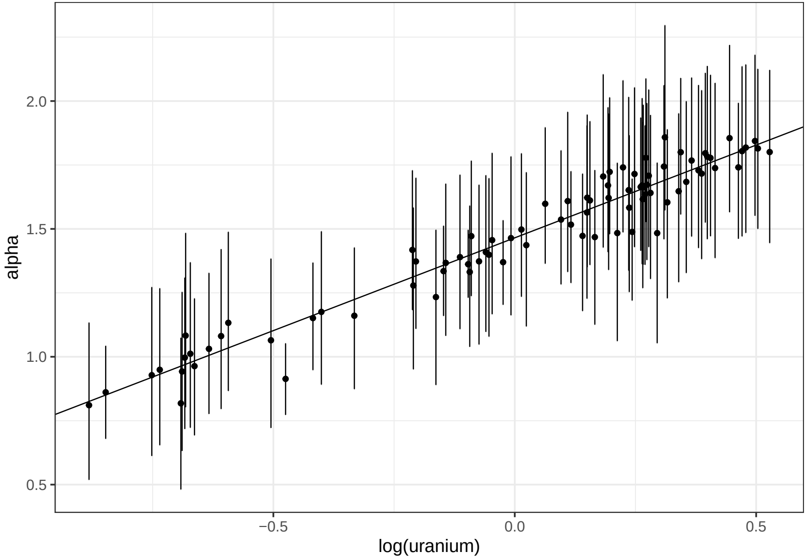
```
model {
  vector[N] y_hat;
  vector[J] alpha_hat;
  for (i in 1:N)
    y_hat[i] = alpha[county[i]] + x[i] * beta;

  for(j in 1:J)
    alpha_hat[j] = gamma0 + gamma1*u[j];

  alpha ~ normal(alpha_hat, sigma_alpha);
  beta ~ normal(0, 1);
  sigma ~ normal(0, 1);
  mu_alpha ~ normal(0, 1);
  sigma_alpha ~ normal(0, 1);

  y ~ normal(y_hat, sigma);
```

Results

```
##                mean se_mean   n_eff Rhat
## beta          -0.66       0 2140.15 1.00
## gamma0         1.47       0  772.98 1.00
## gamma1         0.73       0 1187.67 1.00
## sigma          0.77       0 2703.06 1.00
## sigma_alpha    0.16       0  107.28 1.02
```

- what's the interpretation of beta?
- what's the interpretation of gamma1?

Illustration of model fit $\alpha_j \sim N\left(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2\right)$ for MN



Extending the model: varying slopes

- The last model we discussed for radon included predictors on house and county level
- In that 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 and/or uranium level?
- Focus on set-up for now, fit later.

Extending the model: varying slopes

To start: Focus on floor covariate and leave out the uranium covariate for now.

In this model:

$$y_i|\alpha_{j[i]} \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right), \alpha_j \sim N\left(\mu_\alpha, \sigma_\alpha^2\right)$$

we assume the difference across floors is the same across all houses, regardless of county.

Let's extend:

$$\begin{aligned} y_i|\alpha_{j[i]}, \beta_{j[i]} &\sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right) \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim N_2\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right) \end{aligned}$$

What is : μ_β? σ_β^2? ρ?

Including group-level predictors

- What if the levels and slopes depend on the uranium levels in the county?
- Add back in our group level covariate u_i, where does it go?

Including group-level predictors

$$y_i|\alpha_{j[i]}, \beta_{j[i]} \sim N\left(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2\right)$$

with

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N_2\left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha u_j \\ \gamma_0^\beta + \gamma_1^\beta u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right)$$

- Same as before, but now the mean of the county slopes and intercepts is a function of uranium level
- So we've introduced an interaction between uranium level $u_j[i]$ and floor x_i

Let's rewrite this to see interaction

$$y_i|\alpha_{j|i|}, \beta_{j[i]} \sim N\left(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2\right)$$
$$\left(\begin{array}{c} \alpha_j \\ \beta_j \end{array}\right) \sim N_2\left(\left(\begin{array}{c} \gamma_0^\alpha + \gamma_1^\alpha u_j \\ \gamma_0^\beta + \gamma_1^\beta u_j \end{array}\right), \left(\begin{array}{cc} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{array}\right)\right)$$

write as

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta_{j[i]}x_i + \varepsilon_i \\ \alpha_j &= \gamma_0^\alpha + \gamma_1^\alpha u_j + \eta_j^\alpha \\ \beta_j &= \gamma_0^\beta + \gamma_1^\beta u_j + \eta_j^\beta \end{aligned}$$

with

$$\varepsilon_i \sim N\left(0, \sigma_y^2\right); \left(\begin{array}{c} \eta_j^\alpha \\ \eta_j^\beta \end{array}\right) \sim N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{array}\right)\right)$$

Interactions

$$\begin{aligned} y_i = & \left(\gamma_0^\alpha + \gamma_1^\alpha u_{j[i]} + \eta_{j[i]}^\alpha\right) + \left(\gamma_0^\beta + \gamma_1^\beta u_{j[i]} + \eta_{j[i]}^\beta\right) \cdot x_i + \varepsilon_i \\ = & \gamma_0^\alpha + \gamma_1^\alpha u_{j[i]} + \gamma_0^\beta x_i + \gamma_1^\beta u_{j[i]} x_i (\text{ overall effects }) \\ & + \eta_{j[i]}^\alpha + \eta_{j[i]}^\beta x_i (\text{ county-level effects }) \\ & + \varepsilon_i . \end{aligned}$$

Summary

- Interested in estimating parameters / making inference about a population with a number of groups that naturally form a hierarchy
- Hierarchical models do not estimate group-specific parameters independently of one another but assume a common distribution
- As a consequence, for those groups with much uncertainty about the parameters, estimates are shrunk towards the overall group means.
- Given a multilevel model in math or model output, you should be able to interpret the parameter estimates.
- Given a research problem and relevant data set, you should be able to come up with an appropriate specification of a multilevel model that would provide answers to research questions.
- How to choose between models? you have the tools to do this from last lecture