Hierarchical Bayesian Modeling and Bayesian Model Averaging

DS6040 Fall 2024
Teague R. Henry

SCHOOL of DATA SCIENCE

Outline

- Nested Data - What is it, and why does it matter?
- Hierarchical Bayesian Models in brms
- Ensemble Methods in Machine Learning
- Bayesian Model Averaging - GLMs

Nested Data

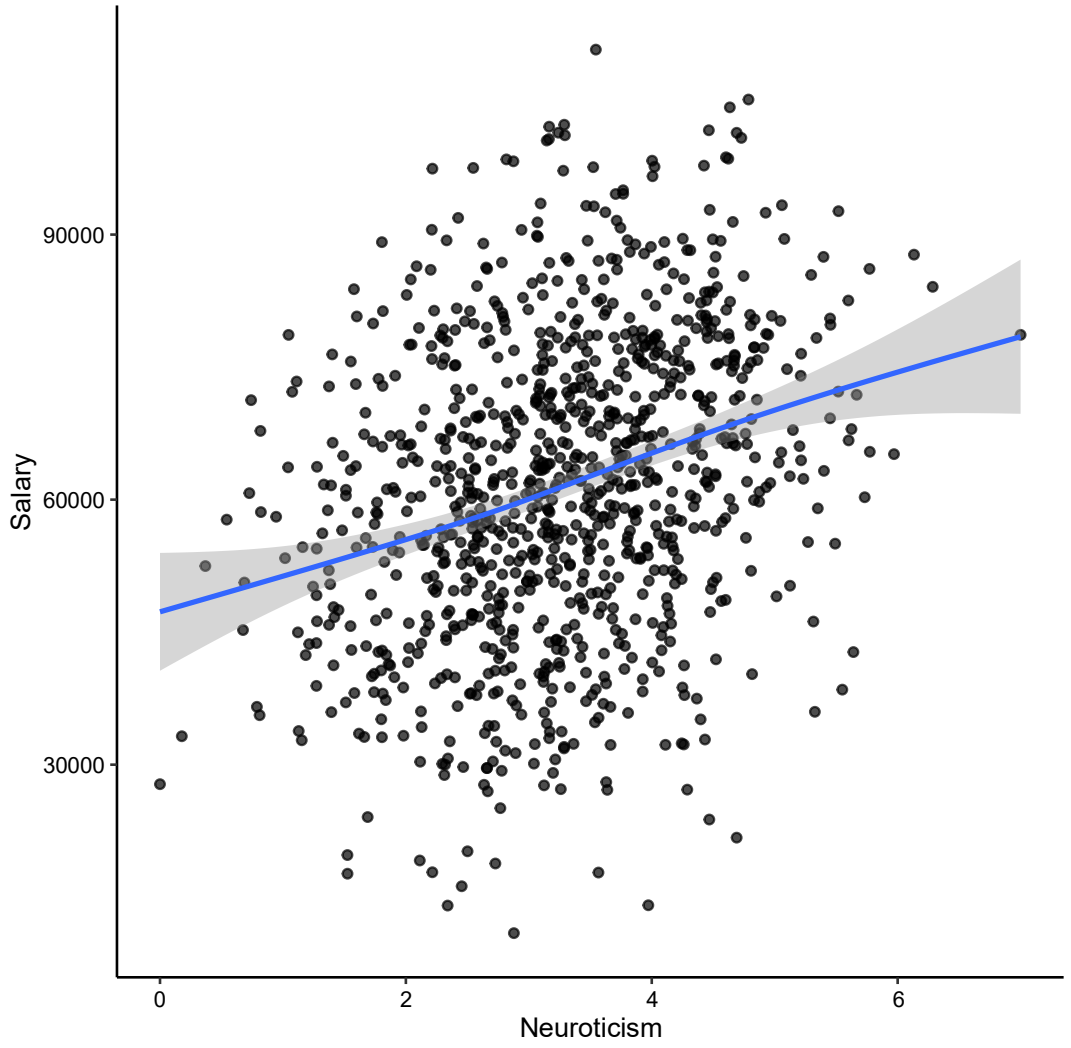
The most used models, Bayesian or Frequentist, assume ***conditional IID***

- Conditional on the model and data, all observations are identically and ***independently*** distributed.

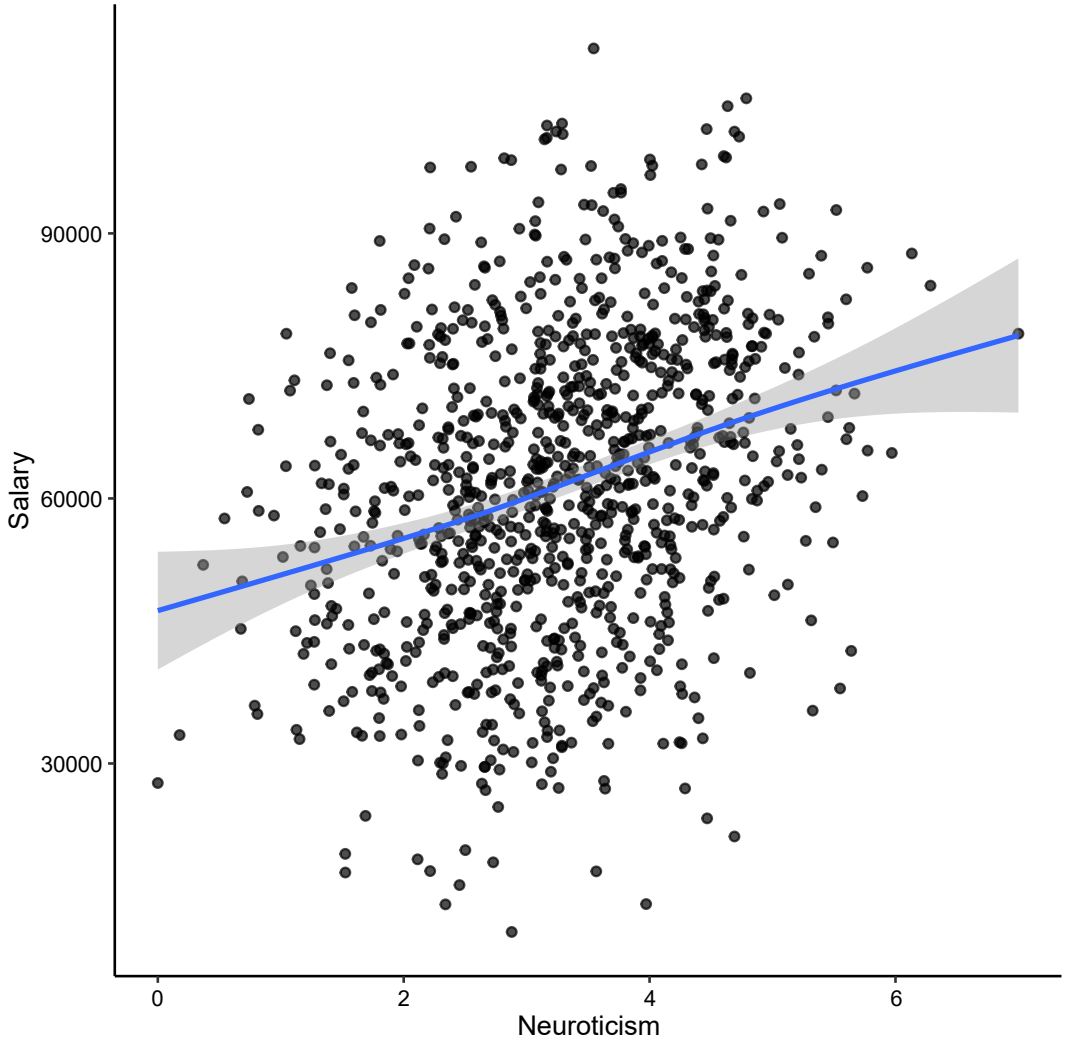
It is rare that any interesting data is conditionally IID.

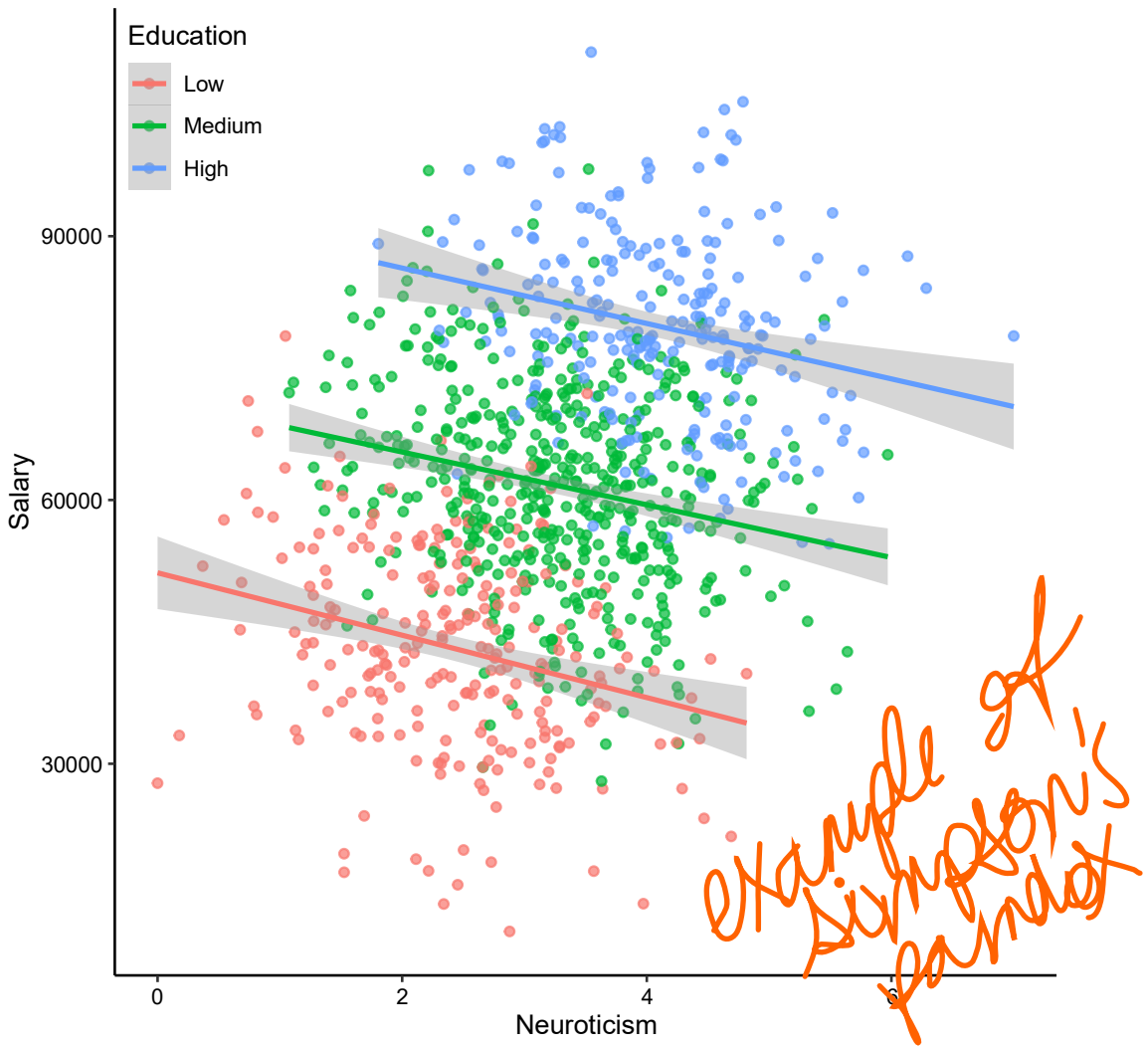
- Sales of big box stores
- Students within classrooms
- Medical history of family members

Violations of conditional IID are **Bad.**



Nested Data





From https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

Simpson's Paradox

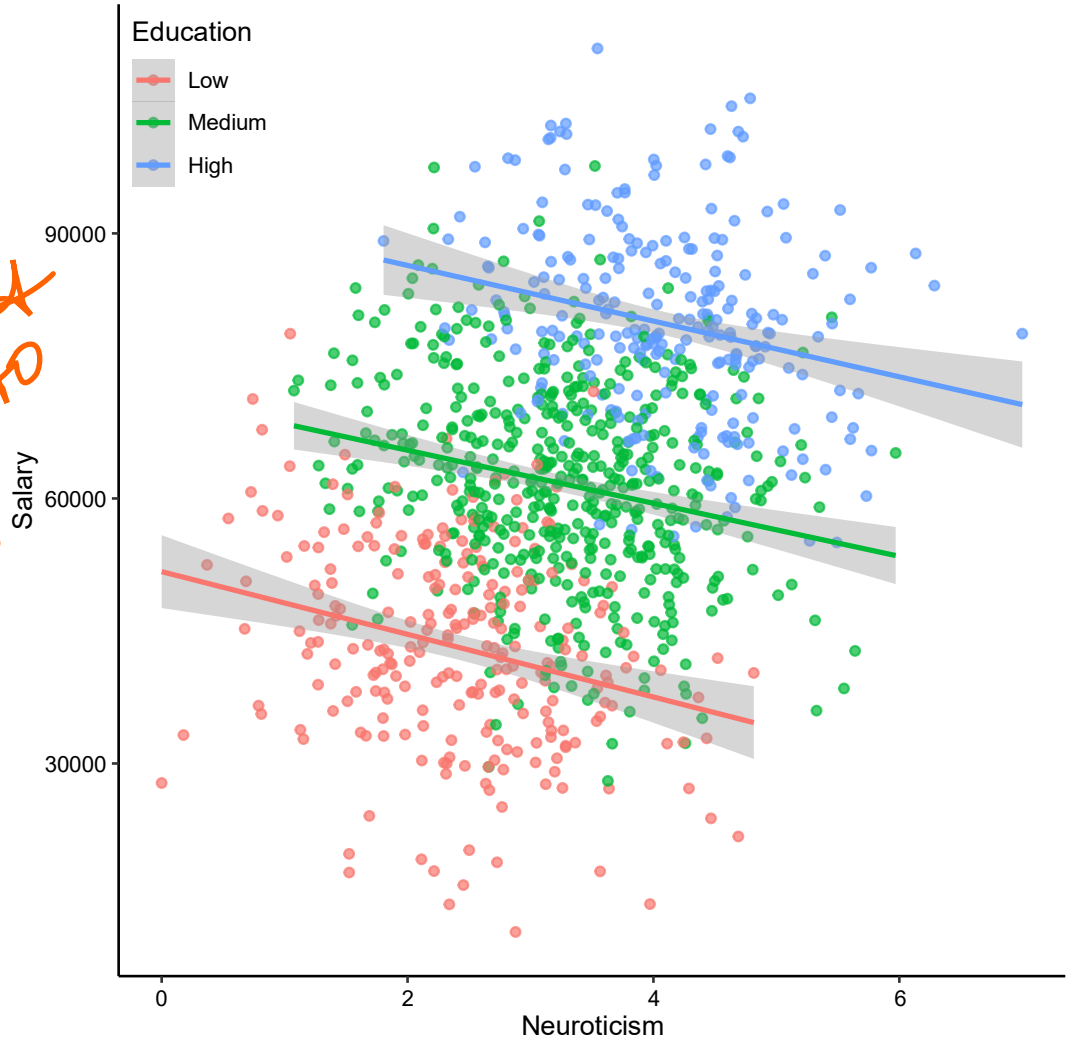
Also known as the aggregation paradox.

- When the overall trend looks nothing like the actual within group trends.
- **Can lead to completely incorrect conclusions.**

Other consequences –

- Inflated standard errors
- Reduction in power
- Very very poor generalizability

might not lead to incorrect predictions



Hierarchical Bayes

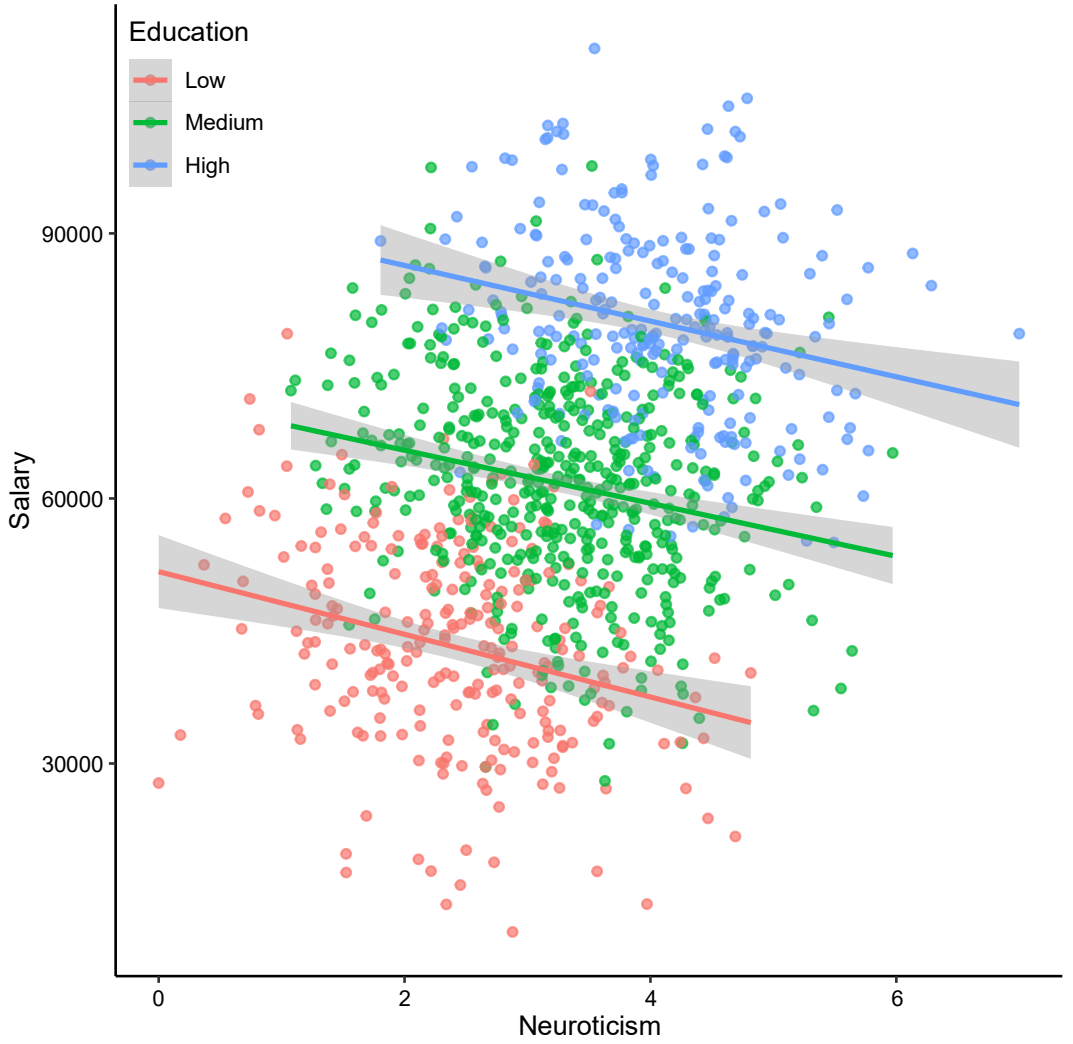
Model the nesting structure by nesting your parameters.

Non-Hierarchical:

$$\beta_0, \beta_1 \sim N(\mu_0, \mu_1, \sigma_0, \sigma_1)$$
$$Salary = \beta_0 + \beta_1 Neuroticism$$

Hierarchical:

$$\beta_0 \sim N(\mu_{education}, \sigma_0)$$
$$\mu_{education} \sim N(\mu_0, \sigma_{00})$$



Hierarchical Bayes Regression

Consider the non-hierarchical regression model:

$$y_i = \beta_0 + X_i\beta + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$
$$\beta, \beta_0 \sim N\left(\mu_\beta, \sigma_\beta^2\right)$$
$$\sigma^2 \sim Half - Cauchy$$

Now, let's allow for nested observations. So, the ith observation of the jth participant are y_{ij} and X_{ij} now.

$y_{ij} = \beta_{0_j} + X_{ij}\beta + \varepsilon_{ij}$

How can we modify this regression equation:

$\beta, \beta_{0_j} \sim \ldots$

1. To allow for differences in β_0 between the j units?
2. To allow for differences in the rest of the βs for j units.

Hierarchical Bayes Regression

$$y_{ij} = \beta_{0j} + X_{ij}\beta + \varepsilon_{ij}$$
$$\beta_{0j} \sim N\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right)$$

Now for the hierarchical priors:

$$\mu_{\beta_0} \sim N(0, 10)\ (Uninformative)$$
$$\sigma^2_{\beta_0} \sim HalfCauchy$$

these look the same as our regular priors, just moved a level up

This is a *random intercept mixed effects model*.

frequentist approach

- Each clustering unit has its own intercept.
- That intercept is drawn from a normal distribution centered at the overall effect μ_{β_0}.
- $\sigma^2_{\beta_0}$ is the variance around that overall effect.

that is, how do people vary from this effect

Hierarchical Bayes Regression

$$y_{ij} = \beta_{0j} + X_{ij}\beta + \varepsilon_{ij}$$
$$\beta_{0j} \sim N\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right)$$
$$\beta \sim MVN(\boldsymbol{\mu}_\beta, \boldsymbol{\sigma}_\beta \boldsymbol{I})$$
$$\mu_\beta \sim N(0, 10)\ (Uninformative)$$
$$\sigma^2_\beta \sim HalfCauchy$$

→ every beta has a person specific mean & sigma.

We can do the exact same thing for the predictor effects. This is called a random slope mixed effects model:

- The effects of each predictor are allowed to vary according to the cluster of the observation.
- This allows us to model things like: The relation between stress and depression differs between people as we model it over time.

Hierarchical Bayes Regression

Key Concepts:

- This use of hierarchical priors allows us to *pool* information across the clustering groups.
  - Regular regression treats the groups as the same, the effects are all set to be equal across the groups.
  - Modeling each group individually would allow for any set of differences, but lowers the available data
  - By posing these hierarchical normal priors, we allow for differences, but constrain them to be normally distributed around the overall effect.
- This general process works for specifying any hierarchical model, not just regression.
- For other response types, all you need to do is swap out the response distribution

Hierarchical Bayes – Radon Example

Radon data set (Gelman et al., 2007)

- Radioactive gas that can seep into houses
- Data from 85 Minnesota counties
- Basements/soil types impact radon
- Great dataset for demonstrating hierarchical Bayes!

Non-Hierarchical Model

$$Radon_{i,c} = \alpha + \beta * floor_{i,c} + \varepsilon$$

$$\alpha, \beta, \varepsilon \sim N(\mu_a, \mu_b, 0, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\varepsilon^2)$$

intercept, error

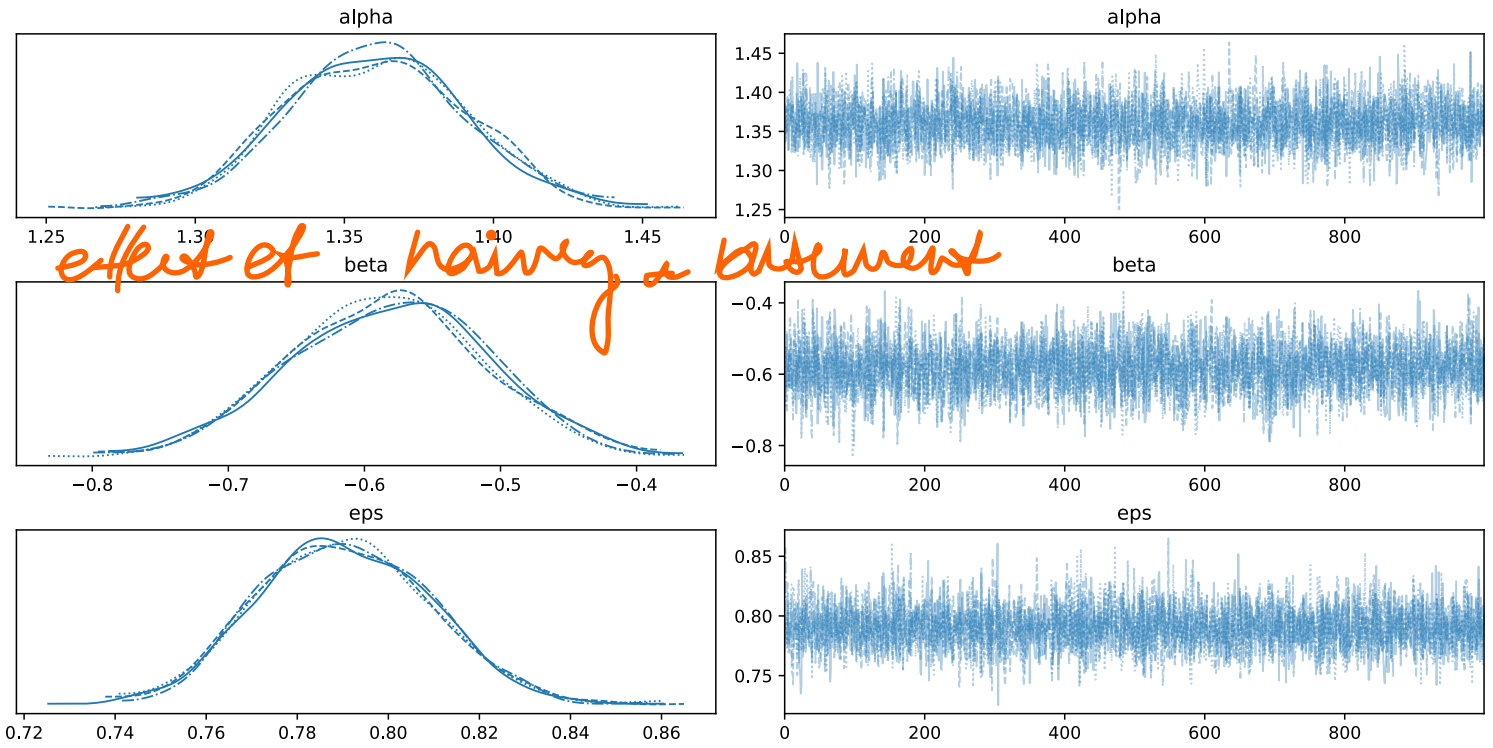
Hierarchical Model

$$Radon_{i,c} = \alpha_c + \beta_c * floor_{i,c} + \varepsilon_c$$

$$\alpha_c, \beta_c, \varepsilon_c \sim N(\mu_a, \mu_b, 0, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\varepsilon^2)$$

This example is a combination of Danne Elbers & Thomas Wiecki's
https://docs.pymc.io/notebooks/GLM-hierarchical.html
And Thomas Wiecki's
https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/

Hierarchical Bayes – Non-Hierarchical



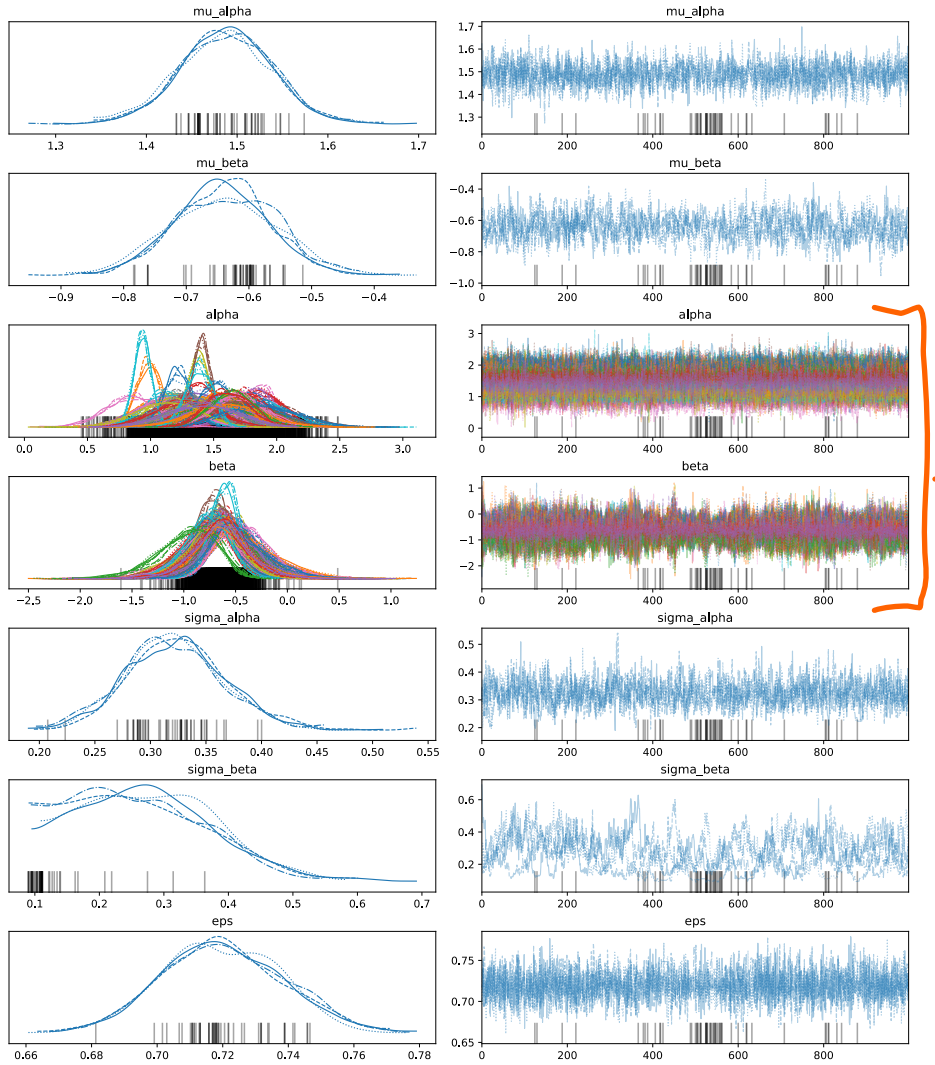
Non-Hierarchical Model

- Note that it converges, it looks fine.
- Nested data doesn't cause convergence issues
- Nested data is a model misspecification issue...

This model looks fine, but it's incorrect.
We find that these effects are no longer representative

This example is a combination of Danne Elbers & Thomas Wiecki's
https://docs.pymc.io/notebooks/GLM-hierarchical.html
And Thomas Wiecki's
https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/

Hierarchical Bayes



Hierarchical Model

- 54 divergences, ESS < 200 for some parameters.
- Substantial variability in the county level intercept, with less in the county level effect of basement.
- Overall, this might seem like it looks good to report…

problems

county specific effects

*The black bars denote divergence events.

This example is a combination of Danne Elbers & Thomas Wiecki's
https://docs.pymc.io/notebooks/GLM-hierarchical.html
And Thomas Wiecki's
https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/

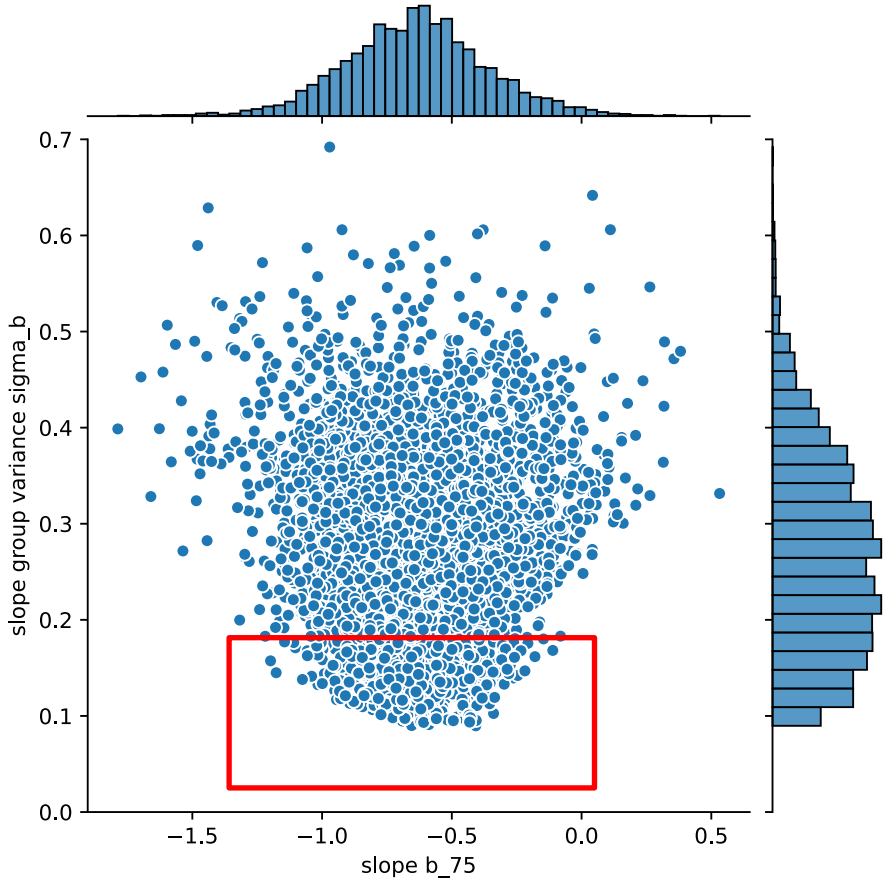
Hierarchical Bayes

Divergences happened at the bottom of the funnel.

- Here, as the variance of β goes to 0, the county level slopes need to converge to the group mean.
- That's a very steep point in the funnel, and the NUTS sampler is having a hard time sampling.

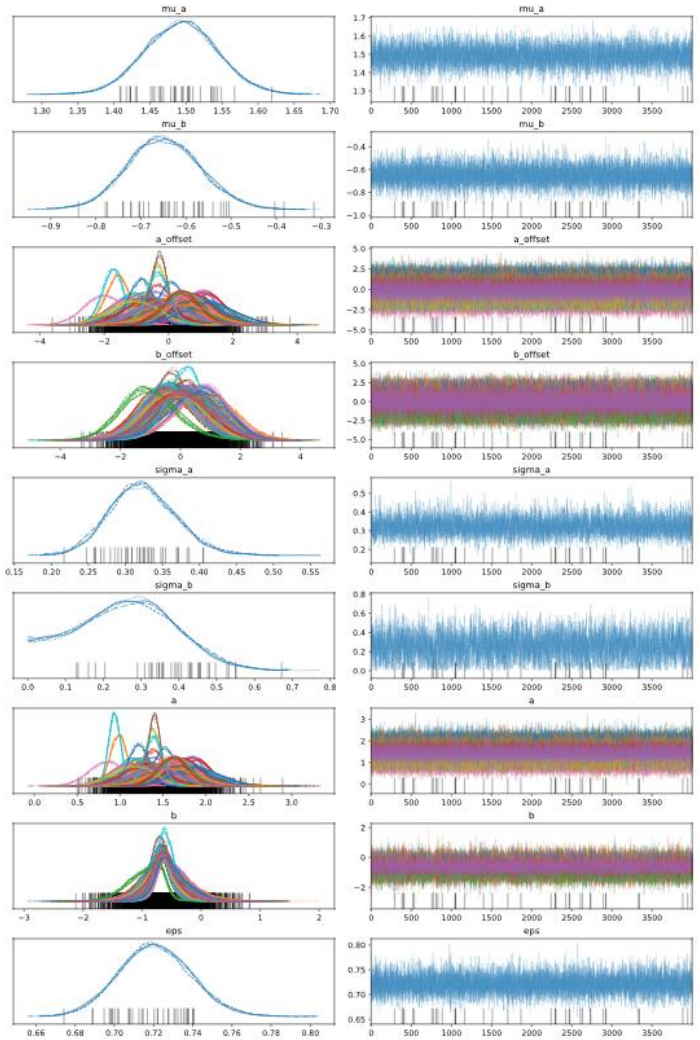
Solution –

- Reparameterize so that slopes are $\beta + \beta_c$
- β is the overall effect of basement
- β_c is the country specific offset



This example is a combination of Danne Elbers & Thomas Wiecki's
https://docs.pymc.io/notebooks/GLM-hierarchical.html
And Thomas Wiecki's
https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/

Hierarchical Bayes- Non-Centered



Fewer divergences, and more uniformly distributed.

- The centered vs non-centered models are statistically equivalent

Hierarchical model let county differences in average radon come through.

- This radically improves our predictive power.

Reparameterization fixed issues with slope estimation

- Usually easier to sample from things with known means. That trick worked here.

This example is a combination of Danne Elbers & Thomas Wiecki's
https://docs.pymc.io/notebooks/GLM-hierarchical.html
And Thomas Wiecki's
https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/

Hierarchical Bayes in brms

Fortunately, you don't need to use Stan directly to quickly put together these models!

Use the `brms` package in R.

- Identical syntax to the lm function, and more specifically to the lme4 package.
- Hierarchical Bayes Regression is literally just mixed effects modeling. All the same caveats and issues apply.

```
library(brms)
library(HLMdiag) #for radon dat
data(radon)
#Radon data has log.radon, uranium, county, basement as variables.
#Basement varies within county, uranium does not.
fit_mod = brm(log.radon ~ basement + uranium + (basement|county), data = radon)
```

implicitly includes random int.

Hierarchical Bayes in brms

```
Group-Level Effects:
~county (Number of levels: 85)
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)            0.13      0.05     0.03     0.25 1.00      984     1233
sd(basement)             0.35      0.12     0.09     0.60 1.00      984     1599
cor(Intercept,basement)  0.18      0.43    -0.66     0.92 1.01     1121     1555

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     1.46      0.04     1.39     1.53 1.00     4176     3193
basement     -0.64      0.09    -0.81    -0.47 1.00     4126     2900
uranium       0.77      0.09     0.58     0.95 1.00     3759     2974

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.75      0.02     0.72     0.79 1.00     4365     2890
```

Hierarchical Bayes in brms

Very easy to fit other types of regressions (logistic, Dirichlet, multinomial)

All you need to do is change the `family` argument in the brm function.

- Default is gaussian, this is standard linear regression.
- For binary outcomes, use bernoulli, binomial, or beta-binomial
  - This is equivalent to logistic regression, when you specify the logit link function.
- Many different types of count distributions, including zero-inflated models.
- The categorical and multinomial families are for multi-response categorical data.
- Models for survival type analyses

All of these families also allow for hierarchical effects!

Ensemble Methods in Machine Learning

Individual models are prone to issues

- Did you choose the right model specification?
- Is the sample truly representative?
- Is the model overfit?

Final Exam is a lot like HW 4.
HW 3 & 4 will use bayes library

One solution is to combine multiple models into an ***ensemble***

- Bagging, Boosting, Random Forests are all ensemble methods

Bayesian Model Averaging was the first ensemble method!

Bayesian Model Averaging

$$P(\Delta|X) = \frac{P(X|\Delta)P(\Delta)}{P(X)}$$

- Let X be the data, which doesn't change across models.
- Δ is a **quantity of interest.** This can be a parameter, or a prediction.

Let $M_1, \ldots, M_k$ be k different models, all with Δ present.

likelihood probability

$$P(M_i|X) = \frac{P(X|M_i)P(M_i)}{\sum_{j=1}^{k} \int P(X|\theta_j, M_j)P(\theta_j|M_j)d\theta_j \, P(M_j)}$$

Where θ_j are the model parameters for model j.

Bayesian Model Averaging

$P(M_i|X)$ is the posterior probability of model i

- Note we are talking fairly abstractly about models here.
- This is often approximated by the Bayesian Information Criteria (BIC), which is itself an approximation of the Bayes Factor

$$BF = \frac{P(X|M_i)P(M_i)}{P(X|M_j)P(M_j)}$$

above 1, evidence is higher for model i
below 1, more evidence for model j

Let $\widehat{\Delta}_i$ be the EAP estimate of Δ from model i. The Bayesian Model Averaged estimate is then:

$$E[\Delta|X] = \sum_{i=1}^{k} \widehat{\Delta}_i \, P(M_i|X)$$

Bayesian Model Averaging

$$E[\Delta|X] = \sum_{i=1}^{k} \widehat{\Delta}_i \, P(M_i|X)$$

What are we doing here?

1. We have a number of models (perhaps different features for a regression)
2. We want to combine our predictions (or parameter estimates), but we also know that some models are going to be better than others.
3. So, using Bayes Theorem twice, we treat the choice of model as a probability distribution, and calculate how likely a given model is given the data.
4. We then use that to weight our predictions or parameter estimates.

Bayesian Model Averaging - Issues

BMA seems simple, but is not that straightforward

- What space of models are you looking over?
  - All posterior probabilities for models are relative to the set of models you are examining.
- You'll need to fit each model individually, are they easy to estimate?
  - You might have 1000s of models to estimate.

Most applications of BMA rely on properties of the general linear model

- i.e. linear, logistic regression
- Taking advantage of conjugate prior distributions.

Word of Warning: BMA is a very complex field, with many super advanced methods that solve important issues (e.g. mixtures of hyper g-priors). Think carefully when applying BMA.

Bayesian Model Averaging- Wine Alcohol

We can use Bayesian Regression to predict red wine alcohol content from just 2 other predictors.

- Question: What are the best predictors of alcohol content?
- Answer: Use BMA to compute variable inclusion probabilities!

Variable Inclusion Probabilities –

- Represents how “often” a variable is in a good model

Variable	Probability of Variable Inclusion
Fixed Acidity	.998
Volatile Acidity	.053
Citric Acid	1
Residual Sugar	.992
Chlorides	1
Free Sulfur Dioxide	.348
Total Sulfur Dioxide	1
Density	.999
pH	1
Sulphates	1

Summary

Hierarchical Models –

- Allow parameters to be random variables with respect to clustering
- Useful for modeling nested data
- Not modeling nested data is problematic

Bayesian Model Averaging –

- Theoretically, any set of models can be averaged together
- Practically, you should really only be averaging regression-type models
- If you'd like to use advanced BMA methods, just be very cautious, there are a number of paradoxes that can occur with BMA…
  - For example – with huge amounts of data, standard BMA selects 1 model and ignores the rest.