Group-level predictors

Patrick Breheny

March 21

Recap

- In our last lecture, we started to look at modeling residential radon levels, with the goal of understanding risk factors for high exposure as well as identifying at-risk homes
- The "identical parameters" approach was clearly unsatisfactory – if the goal of the study is to identify at-risk homes, making an assumption that all counties have the same intercept runs prevents us from answering the question
- On the other hand, the "independent parameters" had problems too: some counties had just a few observations, resulting in highly variable estimates with large standard errors

Group level predictors

- We then fit a hierarchical model, which allowed us to strike a proper balance between the two extremes
- This resulted in a smooth transition between extremes, producing reasonable results for counties of all sizes without any ad-hoc decisions or thresholds
- However, this model assumed that all counties are exchangeable; what if we had county-level information that might be helpful in predicting which counties had higher intercepts than others?
- Our goal for today is to continue looking at the radon data, but add a model at the county level as well as the house level, incorporating what we will call a *group-level predictor*

Soil uranium

- For this data, a potentially relevant county-level predictor is the concentration of uranium in the soil (in parts per million), since radon is one of the decay products of uranium
- Uranium is a naturally occurring element found in low levels within all soils, but can range quite a bit: in our data set, county uranium measurements were taken and ranged from 0.4 ppm to 1.7 ppm
- County-level measurements are posted online in the data set `radon-county.txt`; as before, we will confine our attention to the data from Minnesota

Soil uranium

- Let u denote the log of the soil uranium concentration
- Recall that y was defined as the log of radon levels; thus, an additive model for u and y implies a multiplicative relationship between uranium and radon concentrations (if uranium doubles, we expect a proportional increase in radon levels, say, 30%)

```
AllData2 <- read.delim("radon-county.txt")
Data2 <- subset(AllData2, st=="MN")
u <- log(Data2$Uppm[match(levels(county), Data2$cty)])
```

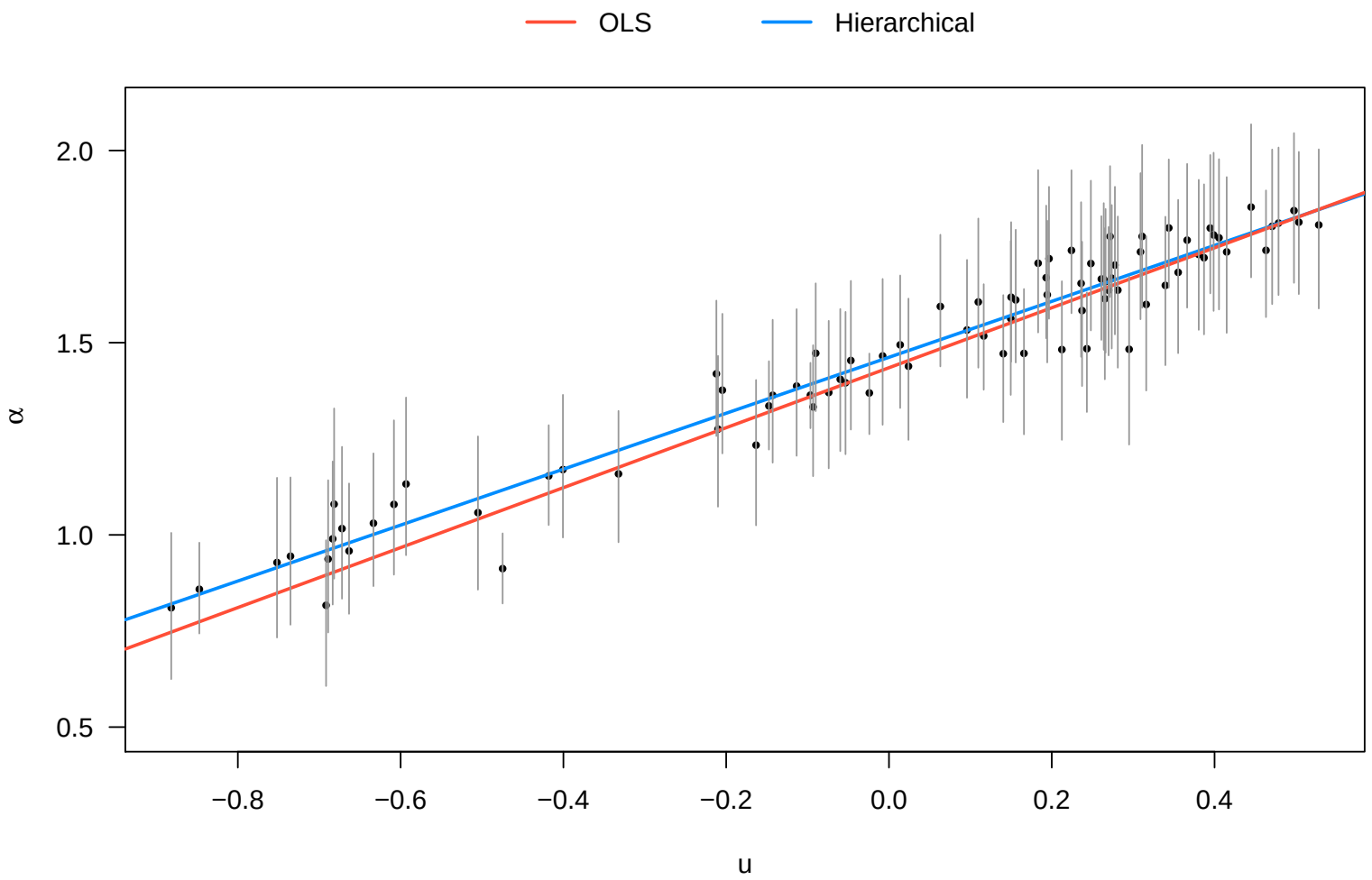
Hierarchical model with group-level predictor

- We now modify our hierarchical model from the previous lecture to include soil uranium concentration as a group-level predictor:

$$Y_{ij} \sim \mathrm{N}(\alpha_j + \beta x_{ij}, \sigma_y^2)$$
$$\alpha_j \sim \mathrm{N}(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2),$$

  with $\boldsymbol{\gamma}$, β, σ_y, and σ_α given uninformative/reference priors
- Note that we now have simple linear regression models at both the county level and house level

County-level model



Posterior for γ_1

- Our posterior for γ_1 has median 0.72, 95% PI: (0.54, 0.90)
- However, our log-log model implies that, if the uranium concentration doubles, the radon concentration increases by 2^{γ_1}; this may be a more relevant and interpretable quantity of interest than γ itself
- Its posterior median is 1.65, with 95% PI: (1.45, 1.87)
- This makes perfect sense: if uranium concentrations double, radon levels should increase, but not double (a 65% increase seems reasonable)

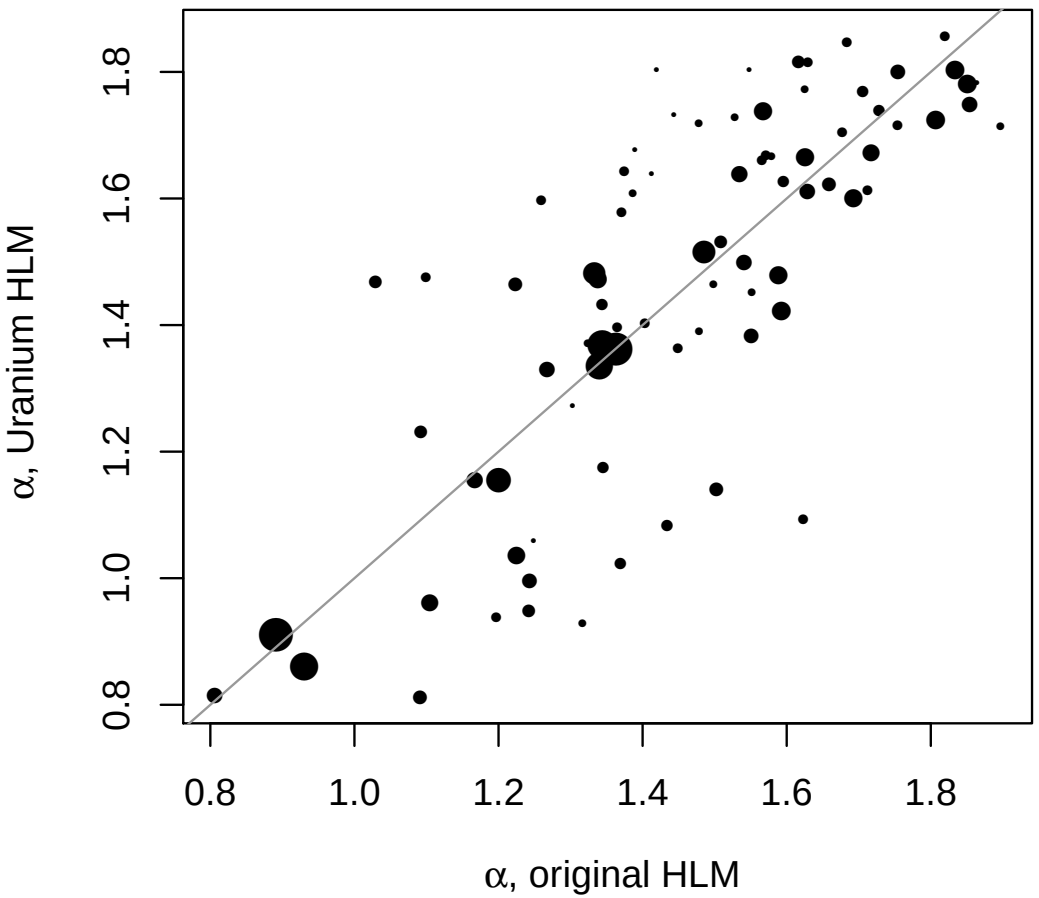
Other parameters

- Our inference regarding β is essentially unaffected:

$$\text{Original HLM} : -0.69(-0.83, -0.56)$$
$$\text{Uranium HLM} : -0.67(-0.80, -0.53)$$

- However, because uranium is so effective at explaining county-level intercepts, σ_α drops from 0.33 to 0.16
- Correspondingly, the intraclass correlation coefficient drops from 0.16 to 0.04

Comparing α, old vs. new models



Moving the uranium term to the house level

- We have presented this model as a combination of two regressions: one on the house level and one on the county level
- However, there are other ways of writing the same model
- For example, the uranium term can be moved to the house level:

$$Y_{ij} \sim \mathrm{N}(\alpha_j + \gamma_1 u_{ij} + \beta x_{ij}, \sigma_y^2)$$
$$\alpha_j \sim \mathrm{N}(\gamma_0, \sigma_\alpha^2),$$

where u_{ij} is simply u_j for all observations in group j

Moving the constant term to the house level

- Of course, we can move the intercept down a level as well:

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 u_i + \beta x_i + \mathbf{z}_i^T \boldsymbol{\gamma}, \sigma_y^2)$$
$$\gamma_j \sim \mathrm{N}(0, \sigma_\alpha^2),$$

  where I am rewriting the model a bit now so that Y, u and x are no longer indexed by j, and $\mathbf{Z}$ is now an $n \times 85$ matrix of county indicators
- This looks a bit different, but is again equivalent to the other two models (with β_0 and β_1 replacing γ_0 and γ_1)
- Note that this is similar to ridge regression, although only the γ parameters are being given what we called "skeptical" priors

One big regression with correlated errors

- Finally, we could write the model as

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where

$$\begin{aligned}
\Sigma_{ii} &= \sigma_y^2 + \sigma_\alpha^2 \\
\Sigma_{ij} &= \sigma_\alpha^2 \qquad i, j \text{ in same group} \\
\Sigma_{ij} &= 0 \qquad i, j \text{ in different groups}
\end{aligned}$$

- Gelman and Hill: "We generally prefer modeling the multilevel effects explicitly rather than burying them as correlations, but once again it is useful to see how the same model can be written in different ways"

Identifiability

- It is worth pausing for a moment here to discuss some practical issues, as we are starting to fit some fairly complicated models
- In particular, let's consider what we did today: we fit a model with 85 intercepts (one for each county), plus a county-level predictor
- Thus, we have 85 counties, and we are attempting to fit 86 county-level parameters; isn't that non-identifiable?

Identifiability (cont'd)

- In a certain sense, indeed they are: you cannot use ordinary least squares to fit this model
- The multilevel modeling is essential here, giving us two synchronized levels on which we do the modeling: the county level has $n \approx 85$ and three parameters, while the house level has $n \approx 919$ and 86 parameters
- However, as we start to flirt with non-identifiability, it is very easy to end up with a model that won't converge, and it can be quite difficult to diagnose the problem

Prior on σ

- For example, consider the seemingly unimportant difference between the following priors on σ:

```
sigma[j] ~ dunif(0, 100)
```

and

```
sigma[j] ~ dgamma(0.001, 0.001)
```

- Both are uninformative and have little impact on the posterior, but the second option mixes much more slowly and takes far longer to converge

Debugging

- This is easy to fix once you know what is causing the problem, of course, but can be extremely difficult to diagnose
- Our textbook has quite a bit of useful practical advice on this topic, especially in Chapters 16.9 and 19
- "In almost any application, a good starting point is to run simple classical models in R and then replicate them in BUGS, checking that the estimates and standard errors are approximately unchanged"

More advice

- "It is usually a mistake in BUGS to program a complicated model all at once; it typically will not run, and then you have to go back to simpler models anyway until you can get the program working"
- It is important to distinguish between mixing that is slow, in that you can just run things a little while longer and it'll be okay, and things that are just too slow (say, even after 100,000 iterations, you only have an effective n under 100)
- "We generally recommend *against* the 'brute force' approach"

Personal experience

In my personal, subjective experience, I would say that (aside from syntax errors) my most common mistakes/remedies in BUGS/JAGS modeling are:

- My priors are too vague
- My initial values are too wild (this tends to be more of an issue with BUGS than JAGS)
- My model is too ambitious, or just not well thought-out