Varying intercepts and slopes

Patrick Breheny

March 26

Varying intercepts and slopes

- Today we take up models in which not only the intercepts, but also the regression coefficients themselves (the slopes) can vary by group
- To put this in terms of our radon example, our previous models assumed that the difference between basements and first-floor radon readings was the same in each county
- We now consider relaxing that assumption, and allowing some counties to perhaps have a greater difference between basements and first floors than others

Radon data

- Of course, not all counties even have first-floor measurements
- In the data set, 25 counties have no first-floor radon measurements (all counties have at least one basement radon measurement)
- Clearly, we cannot learn anything about the difference between first floor and basement radon levels without at least one measurement
- Note, however, that this does not necessarily prevent us from saying anything about those counties, as the hierarchical model allows us to borrow information across counties using the common prior

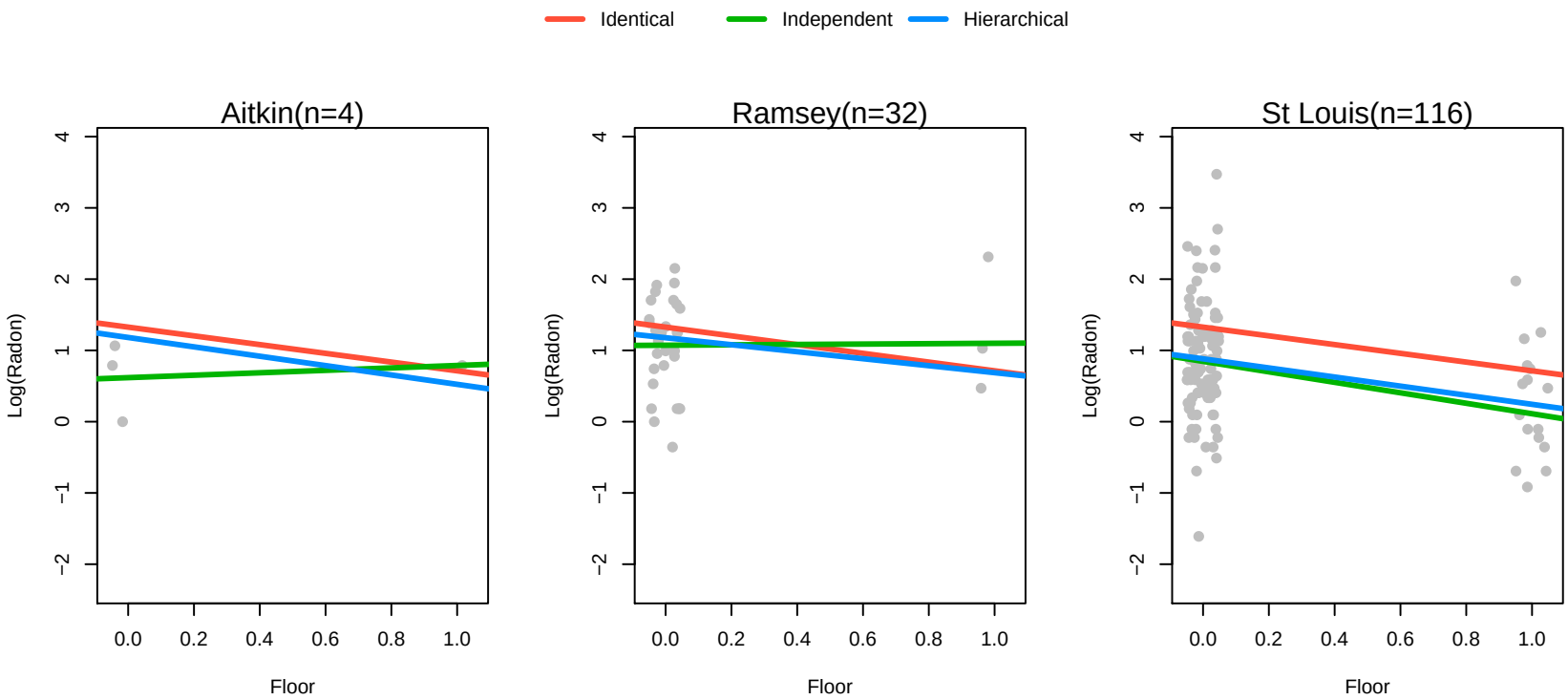
Model #1

Consider the following model:

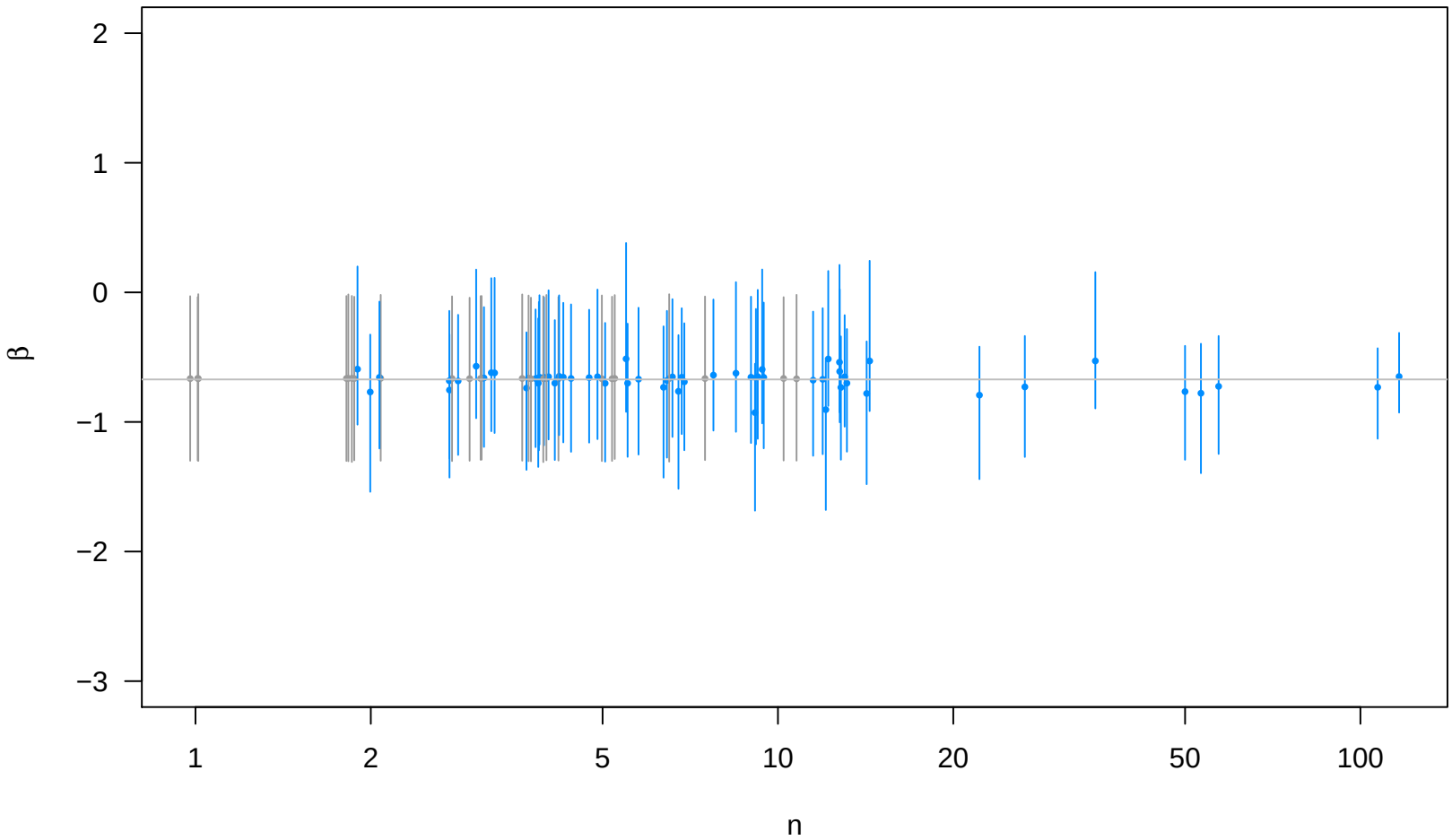
$$\begin{aligned} Y_{ij} &\sim \mathrm{N}(\alpha_j + \beta_j x_{ij}, \sigma_y^2) \\ \alpha_j &\sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2) \\ \beta_j &\sim \mathrm{N}(\mu_\beta, \sigma_\beta^2), \end{aligned}$$

with μ_α, μ_β, σ_α, σ_β, and σ_y given uninformative/reference priors

Regression lines: 3 counties



Posterior intervals for β



Rules of thumb

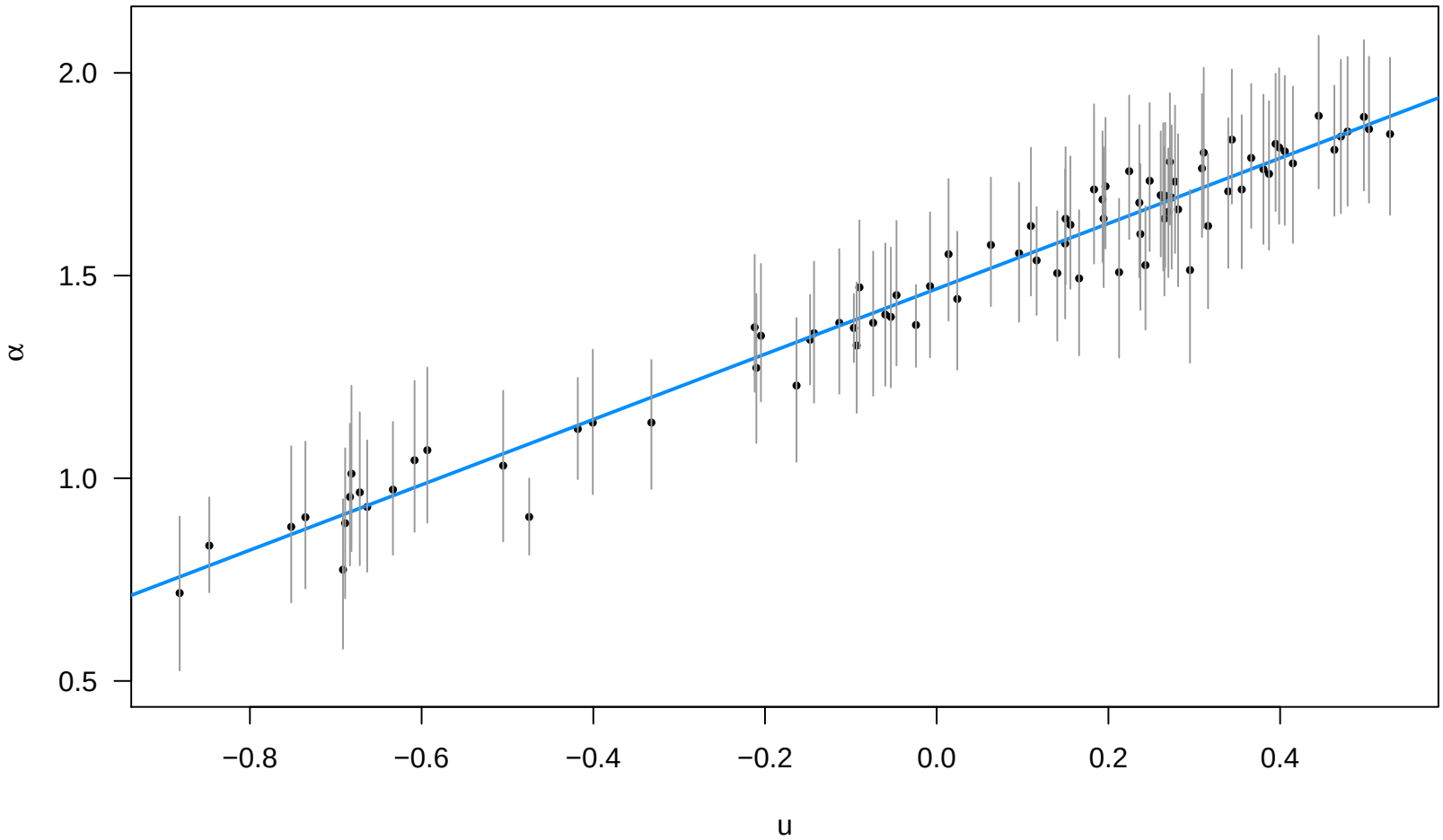
- Gelman & Hill: "Rules of thumb are sometimes given that multilevel models can only be used if the number of groups is higher than some number, or if there is some minimum number of observations per group. Such advice is misguided."
- As the previous model indicates, even when there are effectively zero observations (more precisely, zero information), the multilevel model produces reasonable results
- Of course, for this model, we don't really seem to have gained much by allowing county-level variation in slopes

Model #2

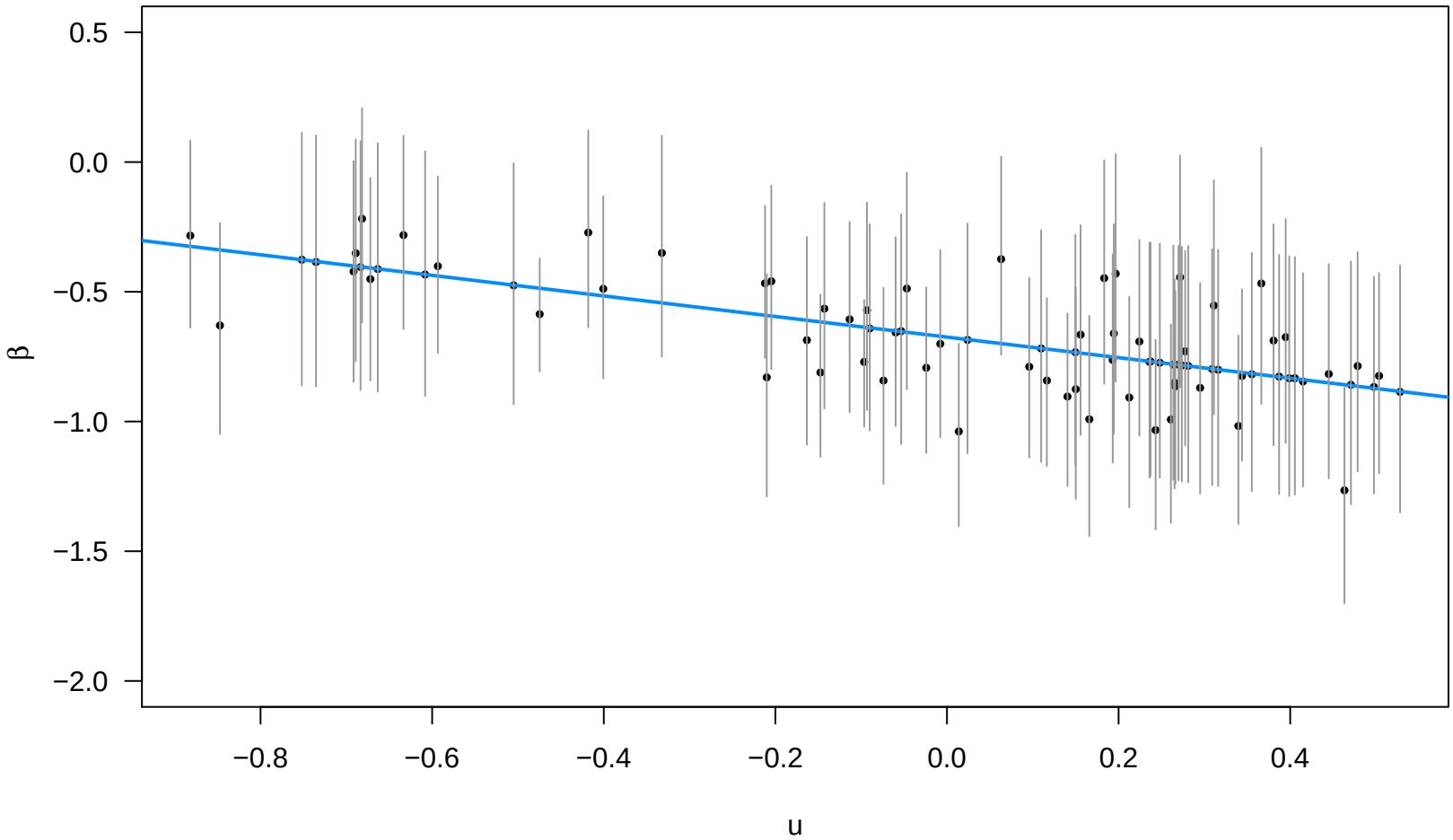
- Recall, however, that we had a rather useful county-level predictor for radon levels: soil uranium measurements
- An improvement on Model #1 would be to model the county-level intercepts and slopes using soil uranium:

$$
\begin{aligned}
Y_{ij} &\sim \mathrm{N}(\alpha_j + \beta_j x_{ij}, \sigma_y^2) \\
\alpha_j &\sim \mathrm{N}(\gamma_{\alpha 1} + \gamma_{\alpha 2} u_j, \sigma_\alpha^2) \\
\beta_j &\sim \mathrm{N}(\gamma_{\beta 1} + \gamma_{\beta 2} u_j, \sigma_\beta^2),
\end{aligned}
$$

County-level model: α



County-level model: β



Posteriors for γ

- For this model, the 95% posterior interval for $\gamma_{\beta 2}$ is (-0.85, 0.07)
- This is moderately convincing evidence that the difference between basement and first-floor radon measurements is larger in counties with higher soil uranium concentrations
- On the other hand, the 95% posterior interval for $\gamma_{\alpha 2}$ is (0.61, 0.99); we can be certain that average basement radon levels are higher in counties with higher soil uranium concentrations

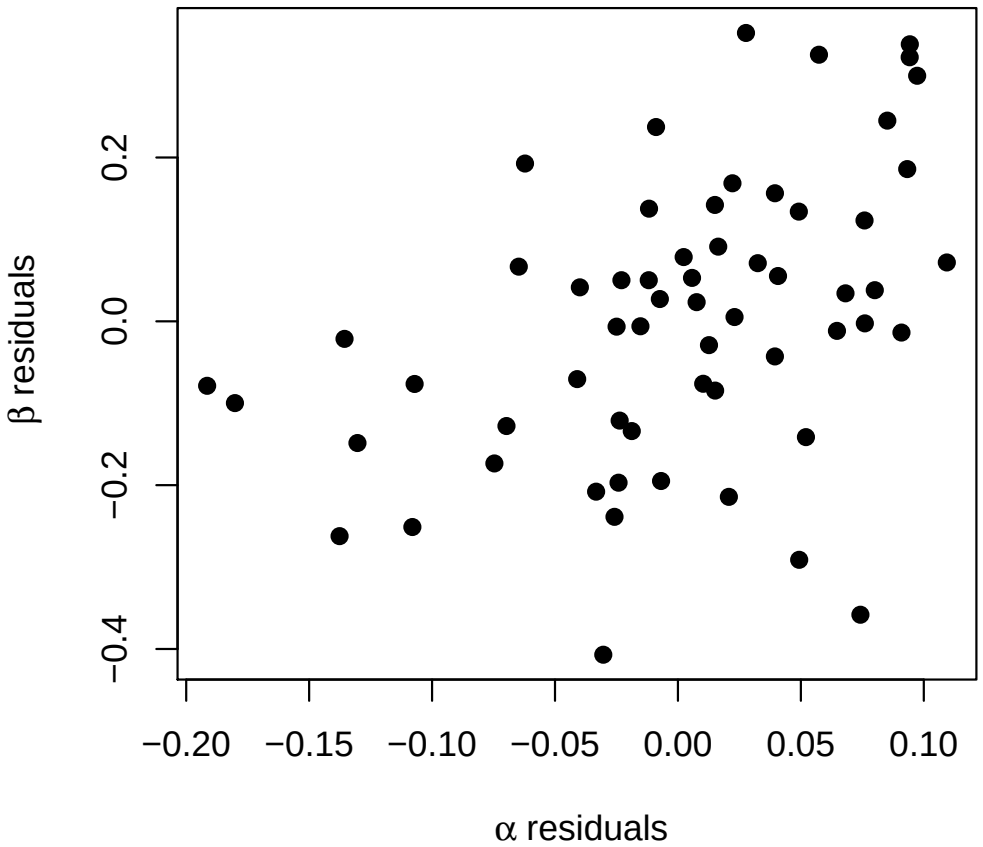
Residuals

Let's examine what we might call the "residuals" of the county-level model:

$$r_{\alpha j} = \bar{\alpha}_j - \bar{\gamma}_{\alpha 1} - \bar{\gamma}_{\alpha 2} u_j$$
$$r_{\beta j} = \bar{\beta}_j - \bar{\gamma}_{\beta 1} - \bar{\gamma}_{\beta 2} u_j,$$

i.e., the difference between the (posterior means for the) county-level intercept/slopes and the levels we would predict based on that county's uranium concentrations

Residuals (cont'd)



Hierarchical correlations

- The correlation between the residuals is 0.4
- This is a common occurrence in hierarchical modeling – group-level parameters are typically correlated
- Our previous models, which assumed independence between α_j and β_j, were instructive building blocks, but in practice, it is usually appropriate to allow correlation

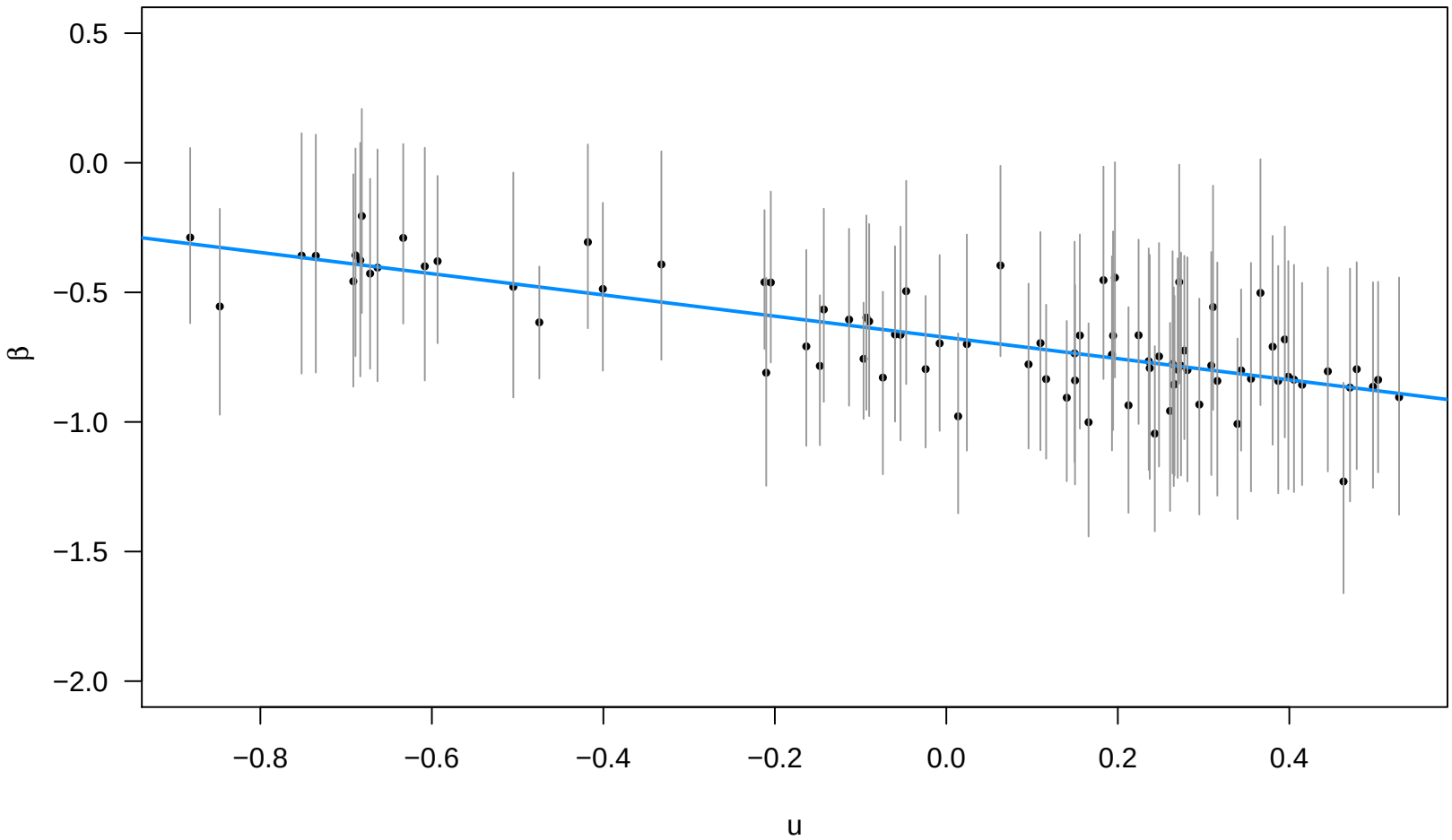
Model #3

Let's consider now model #3, where we allow county-level correlation between intercept and slope:

$$
\begin{aligned}
Y_{ij} &\sim \mathrm{N}(\alpha_j + \beta_j x_{ij}, \sigma_y^2) \\
\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} \gamma_{\alpha 1} + \gamma_{\alpha 2} u_j \\ \gamma_{\beta 1} + \gamma_{\beta 2} u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right) \\
\rho &\sim \mathrm{Unif}(-1, 1),
\end{aligned}
$$

where the γ and σ parameters have the usual uninformative priors

County-level model with correlated parameters



Posteriors for γ, ρ

- In this example, allowing for correlation between α and β does not dramatically affect the results
- In particular, the posterior median for $\gamma_{\beta 2}$ is now -0.41, with a 95% interval of (-0.87, 0.05), very similar to our earlier results
- In part, this arises from the fact that there is limited information about β in many counties, and consequently, limited information about ρ
- The posterior median of ρ is 0.28, with a 95% interval of (-0.62, 0.96)

Posterior draws of the regression line