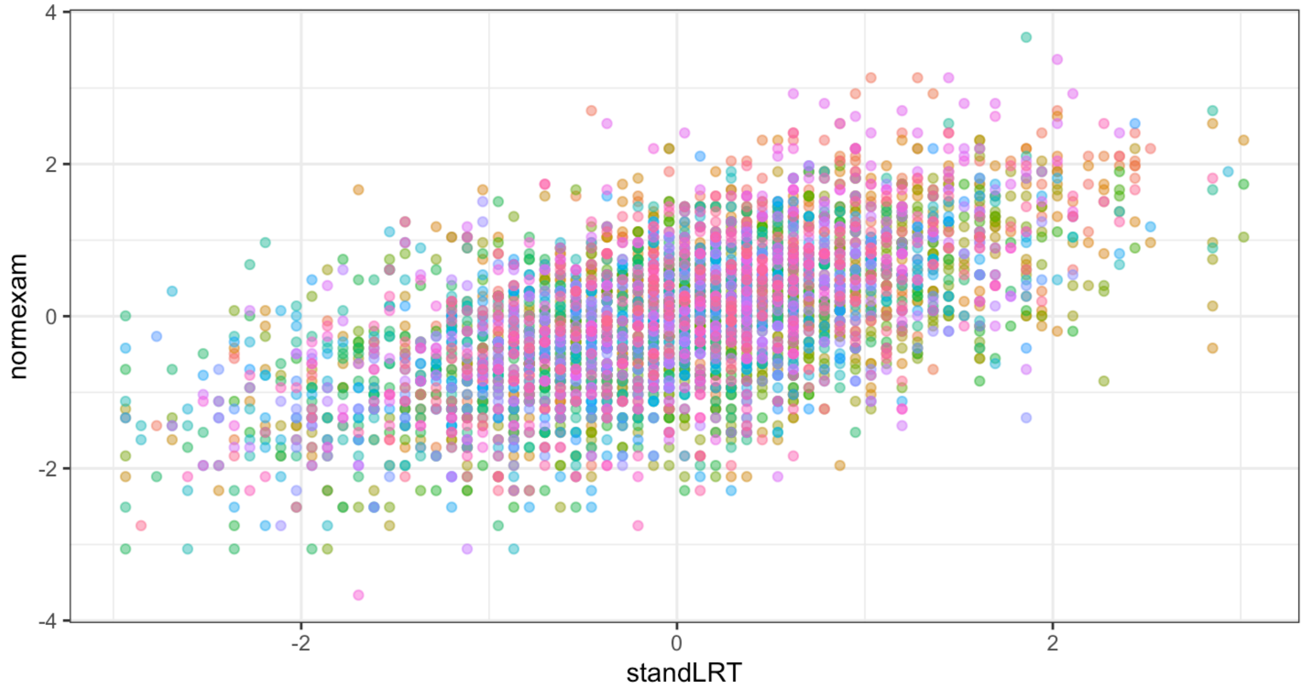
Hierarchical Linear Models

Bayesian Data Analysis

Steve Buyske

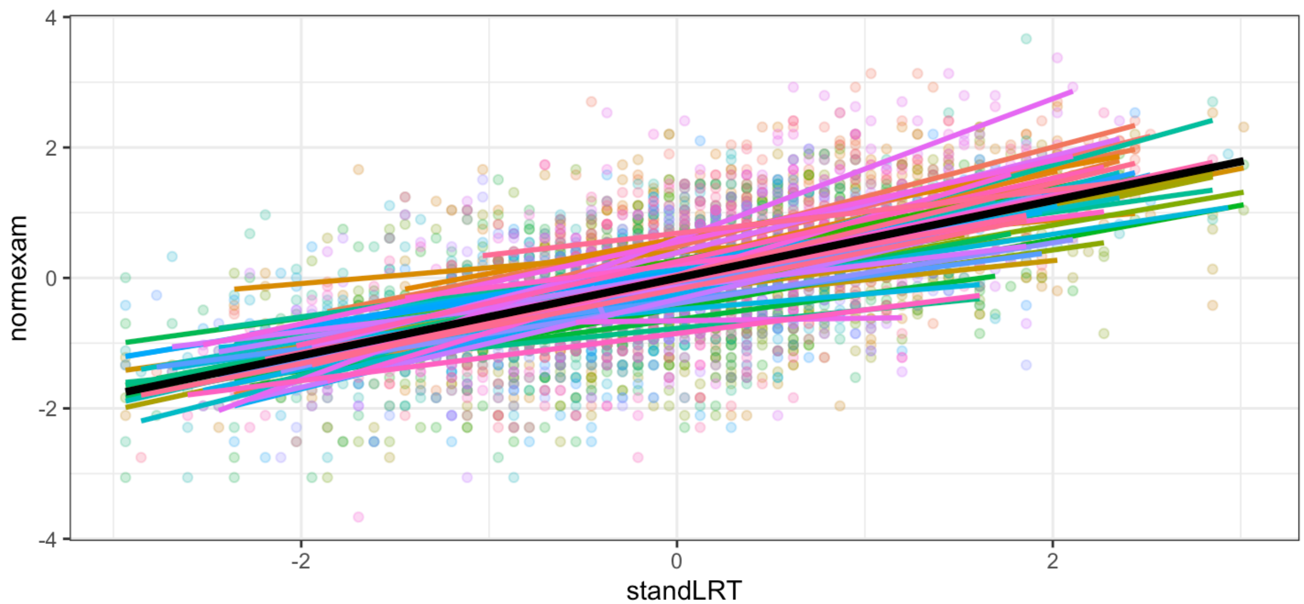
An Example

- Our first data set consists of exam scores at the start and end of the school year for 4059 students. the exams are standardized to mean 0 and standard deviation 1. The plot shows a scatter plot, with the individual observations colored by the 65 schools.



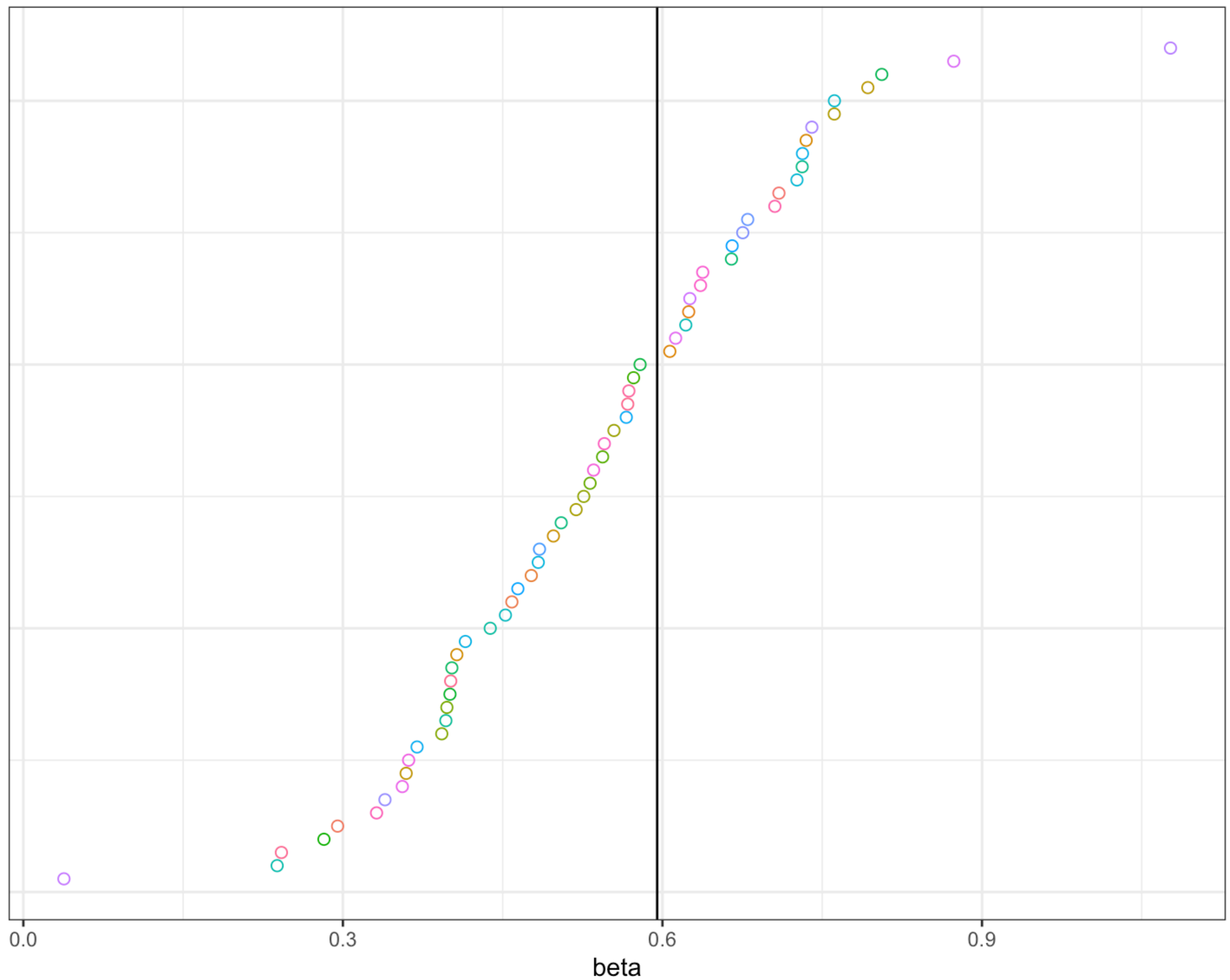
An Example cont.

- Now I've added a least squares line for each school.
- The thick black line is the least squares fit for all of the students at once.
- Clearly there's a lot of variability in the slope (and intercept) by school.



An Example cont.

- The next slide shows a plot of just the slopes.
    - the vertical black line indicates the predicted value if we fit a single regression line to all the students—that is, pool all the data.
- Again, it’s clear that there is a considerable variation from school to school.
- Some of that variation is presumably due to genuine differences, while others represents random variation (aka noise).


0.0
0.3
0.6
0.9
beta

Linear Models When the Data is Grouped

- While it would be legitimate to analyze each school separately, it feels like if we do so we would not be taking advantage of everything we know.
- By looking at all of the slopes at once, we get a sense of the distribution of the slopes.
- We might also conclude that the most extreme slopes are overly extreme estimates, just due to chance.
- That's particularly true if the sample size for that school is small.
  - For example, the two smallest slope correspond to schools of size $n = 2$ and $n = 8$.

Linear Models When the Data is Grouped cont.

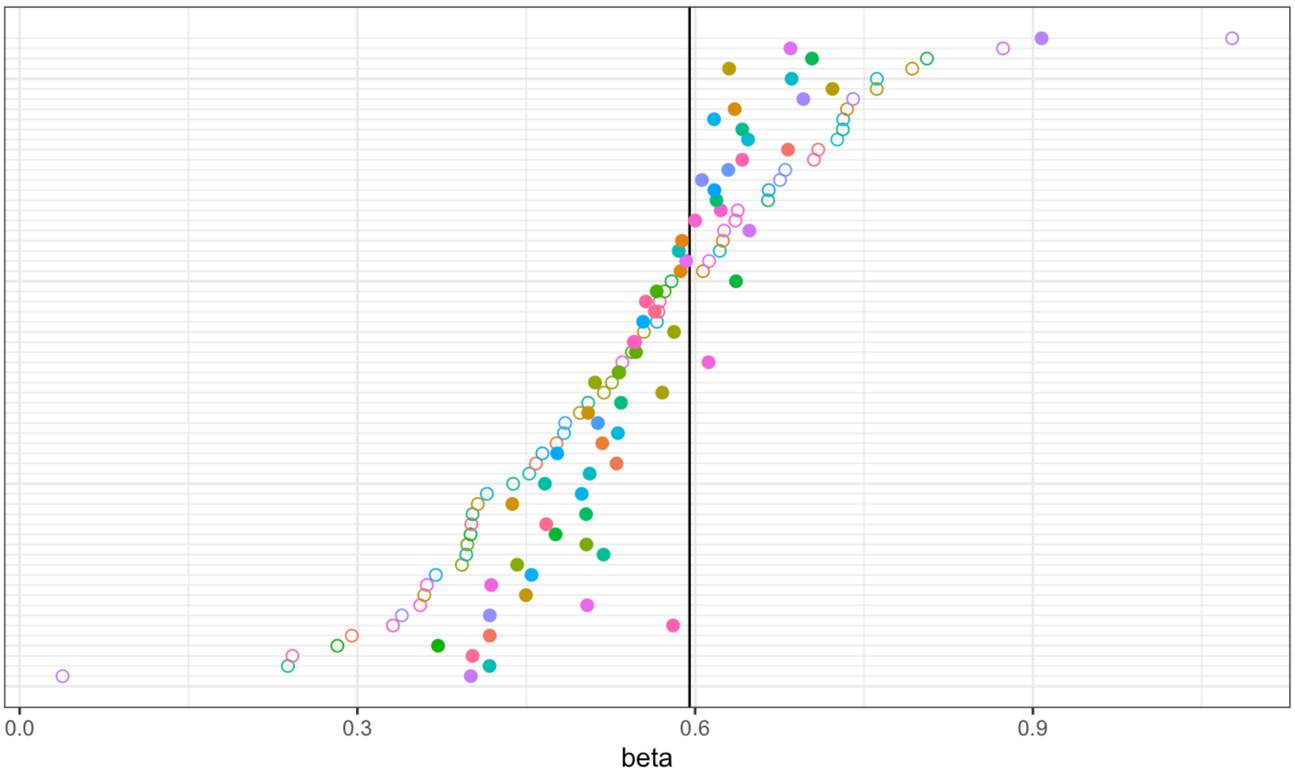
- There are three general approaches one could take when we have this kind of grouped data.
- One approach is to fit each school blind to all the other schools; that is, fit each school separately.
  - Doesn't take into account useful information.
  - Because of the smaller sample sizes, there is a risk of overfitting each regression line.
  - We will refer to this approach as "no pooling" of the data.

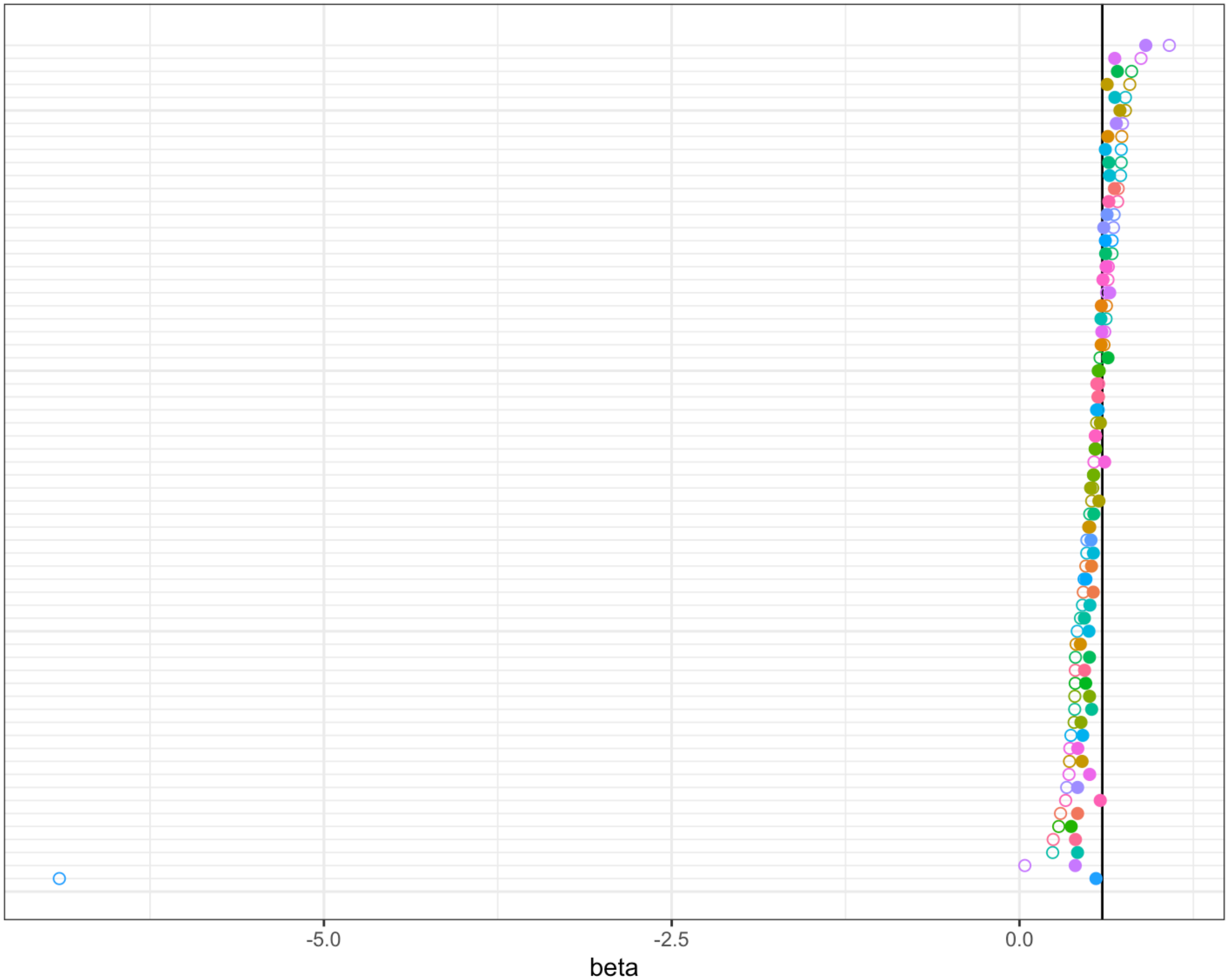
- A second approach is to fit a single regression line; that is fit all students at once while ignoring the group structure.
    - This ignores the school entirely.
    - This risks underfitting the regression line.
        - Ignores potentially important information (the school) and possible correlation due to that.
    - We will refer to this approach as “complete pooling” of the data.

Hierarchical Linear Models

- The third approach is to *partially* pool the the data.
  - That is, conceptually fit each school individually, while taking advantage of our knowledge of the fit of other schools
  - Phrased differently, we can start with an overall regression line, but simultaneously allow for school-specific variation in regression lines.
- This idea has been called “borrowing strength”—our estimates for one school borrow strength from other schools since we know what their estimates look like.
- Another phrase for the phenomenon is called *shrinkage* of the estimates closer together.
- The statistical models are called *hierarchical linear models*, *multilevel models*, or *mixed effects models*.
- Before we look at it formally, this plot on the next slide shows the result of such a model.

- The open circles correspond to the unpooled model.
- The filled circles corresponds to the hierarchical model.
- Notice how the filled circles are generally pulled towards the black vertical line.



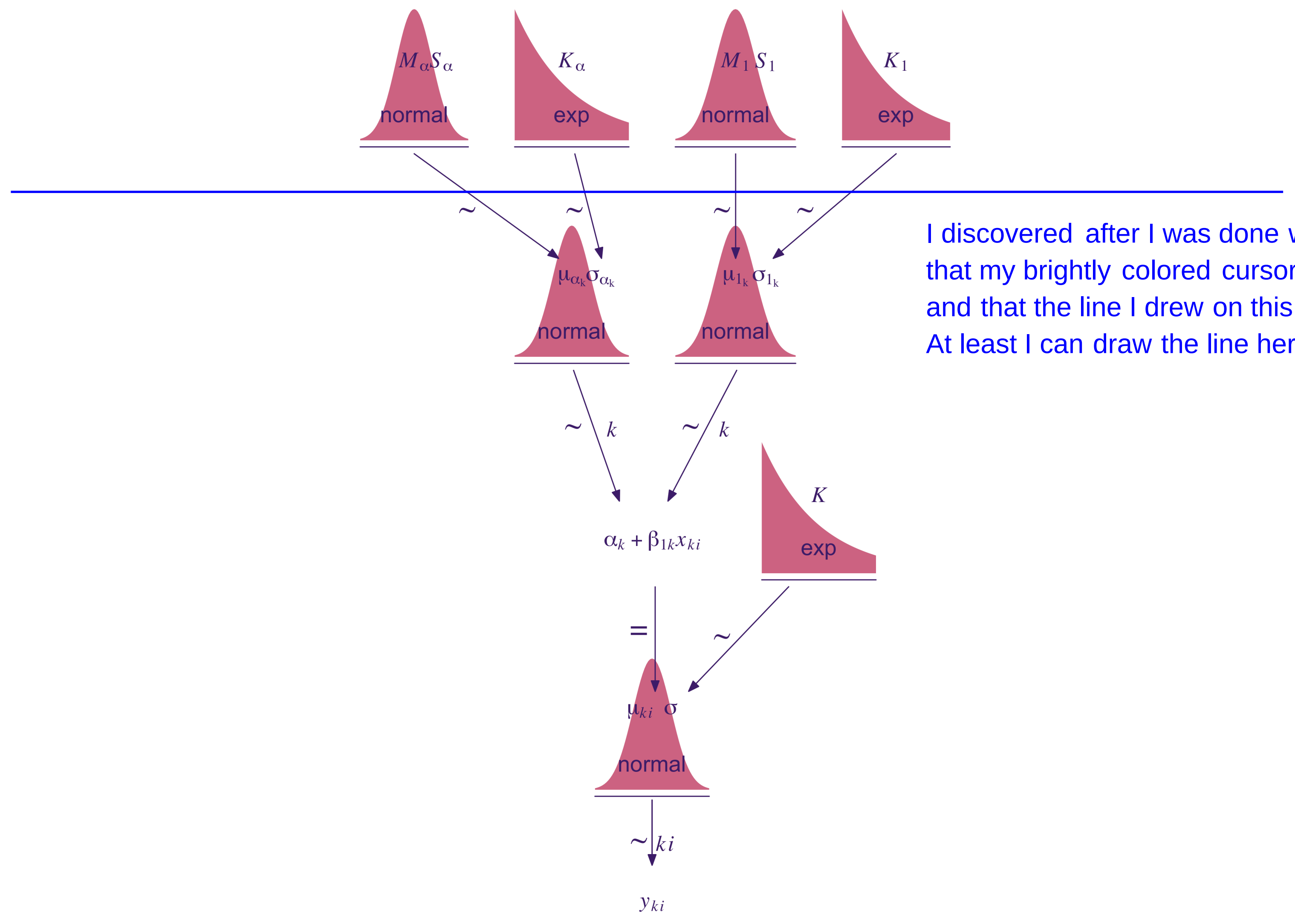

-5.0
-2.5
0.0
beta

Formal framework

- Let's say the data looks like (x_{ki}, y_{ki}), where k indexes the group (the school in our example) and i indexes the individual.
- Each group has a separate intercept and slope.
  - This part looks just like an interaction model
  - It's the relationship among those intercepts and slopes that different.

$$y_{ki} \sim \text{Normal}(\mu_{ki}, \sigma)$$

$$\mu_{ki} = \alpha_k + \beta_{1k} x_{ki}$$



I discovered after I was done with the video
that my brightly colored cursor didn't show up,
and that the line I drew on this slide wasn't visible.
At least I can draw the line here.

Priors for this model

$$
\begin{aligned}
\alpha_k &\sim \text{Normal}(\mu_{\alpha_k}, \text{something}) \\
\mu_{\alpha_k} &\sim \text{Normal}(\text{something}, \text{something}) \\
\beta_{1k} &\sim \text{Normal}(\mu_{1_k}, \text{something}) \\
\mu_{1_k} &\sim \text{Normal}(\text{something}, \text{something}) \\
\sigma &\sim \text{Exponential}(\text{something}).
\end{aligned}
$$

An alternative parametrization of the model

- A more common version of the model is to think of an overall intercept and slope with variation from that.
- In this case, the overall intercept and slope are called the *fixed effects* and the group-specific differences the *random effects*.

$$y_{ki} \sim \text{Normal}(\mu_{ki}, \sigma)$$

$$\mu_{ki} = (\alpha + \gamma_k) + (\beta + \delta_k)x_{ki}$$

Priors for the alternative parametrization

$$\alpha \sim \text{Normal}(\text{something}, \text{something})$$

$$\gamma_k \sim \text{Normal}(0, \text{something})$$

$$\beta \sim \text{Normal}(0, \text{something})$$

$$\delta_k \sim \text{Normal}(0, \text{something})$$

$$\sigma \sim \text{Exponential}(\text{something}).$$

How the hierarchical model works

- Individuals in the same group have the same distribution for α_k and β_k.
- Those parameters **themselves** have distributions
  - That means that extreme values for those parameters, for some specific group, is possible, but less likely than values closer to the modes.
  - That means that the distributions for α_k and β_k for a specific group are pushed towards the middle (aka “shrinkage”).
- How much? …

How much shrinkage occurs?

- The amount of shrinkage is determined by the model, especially the priors, and the data.
    - There's no manual shrinkage that you have to decide on after the analysis.
- In a group with a large sample size, the evidence will get most of the weight in determining α_k and β_k.
- In a group with a small sample size, there will be less evidence about the values of α_k and β_k, so the distributions (higher in the Kruschke diagram) for the mean and sd of *the distributions* for α_k and β_k get more relative weight.
- If there are many groups with similar distributions for the parameters,
    - then the posteriors for the distributions at the top of the diagram will be fairly narrow
    - which will lead to more shrinkage for a group that's atypical.

Summary about grouped data

- Ignore the groups => a single model => “complete pooling” => likely to underfit.
- Include the groups, while fitting everything at once in a hierarchical model => a single model that includes the group effect => “partial pooling” => likely to fit just right.
- Fit each group separately, ignoring the other groups => separate model for each group => “no pooling” => likely to overfit.

- There are details of analyzing the exam data in the RStudio Cloud project.