# Hierarchical Bayesian Linear Regression

June 8, 2021

# Motivation

# Setting

Suppose we collect multiple measurements (covariates & outcomes) within each of several groups.

**Example**

- **Outcome:** math score
- **Covariate:** socioeconomic status (SES)
- **Observational units:** 10th graders
- **Groups:** 100 public high schools

# Separate regression lines

What if we fit a separate regression for each school?

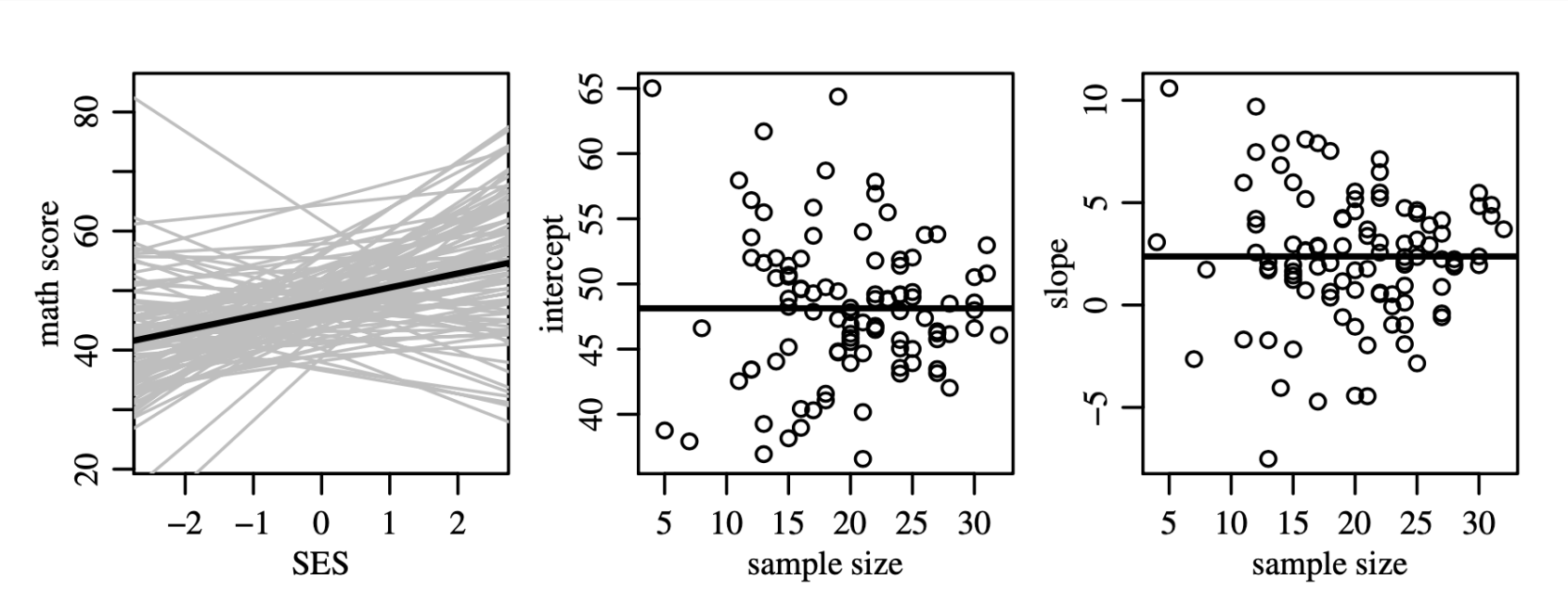



**Figure 1:** Least squares regression lines; estimates vs. group sample sizes

*Observations?*

# Separate regression lines

What if we fit a separate regression for each school?

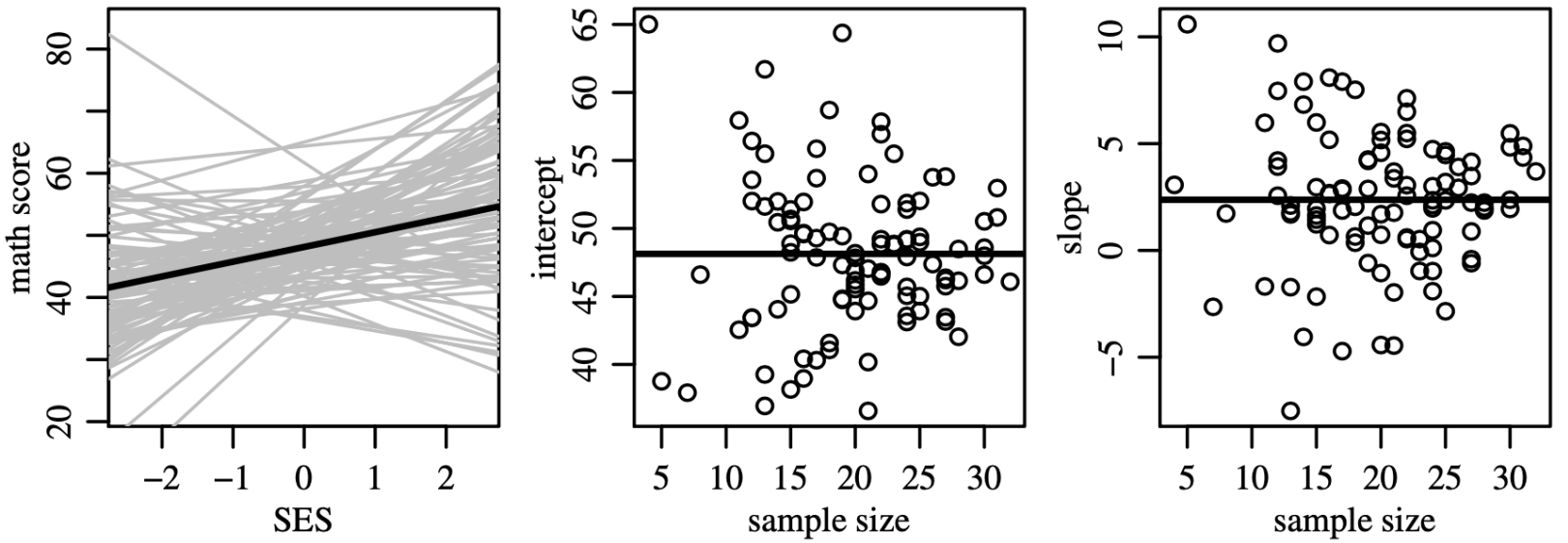



**Figure 1:** Least squares regression lines; estimates vs. group sample sizes

*Observations?* Schools with ...

- ... the highest sample sizes tend to have regression coefficients that are close to the average.
- ... the lowest sample sizes tend to have regression coefficients that are more extreme.

Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

# Problem and solution

## The problem

Regression estimates are unstable when information is low.

Small Datasets

*Example:* Some groups can have few observations

Large Datasets

*Example:*

# Problem and solution

## The problem

Regression estimates are unstable when information is low.

Small Datasets

*Example:* Some groups can have few observations

Large Datasets

*Example:* There exists a rare, but highly predictive, binary covariate. (The estimation becomes especially hard if there are many other covariates, some of them also highly predictive, and some of them correlated with the rare binary covariate.)

## The remedy

Stabilize the regression estimates by sharing information across groups, using a hierarchical model.

# Model

# A model

Consider a Bayesian hierarchical linear regression.

$$
\begin{aligned}
\boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{V}_0) \\
\Sigma &\sim \mathcal{W}^{-1}(\eta_0, \Psi_0) \\
\boldsymbol{\beta}_j &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\
\sigma^2 &\sim \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2) \\
y_{ij} &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\beta}_j^T \boldsymbol{x}_{ij}, \sigma^2)
\end{aligned} \tag{1}
$$

## The idea

We take the regression to be hierarchical in the sense that we take the regression weights $\boldsymbol{\beta}_j$ to be distinct for each of $j = 1, \ldots, J$ groups, but we assume that the $\boldsymbol{\beta}_j$'s are drawn from some distribution. The model allows for "sharing statistical strength" in the sense that uncertainty about the $j$th group's regression parameters, to the extent that it exists, can be reduced by borrowing information from the other groups $k \neq j$. In other words, for grouped data, we allow the information from the other groups to play the role that is played by the prior in Bayesian linear regression. (*Explain.*)

# Inference

# Inference on "Population-level" (i.e. across-group) quantities

Key Insight: The top part of the model behaves like a Bayesian multivariate normal model, but where the "data" are the (latent) regression weights, $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$.

$$\boldsymbol{\mu} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \Sigma \sim \mathcal{N}(\boldsymbol{m}', \boldsymbol{V}')$$

$$\boldsymbol{m}' = \boldsymbol{V}' \left( \boldsymbol{V}_0^{-1} \boldsymbol{m}_0 + J \, \Sigma^{-1} \, \overline{\boldsymbol{\beta}} \right), \qquad \overline{\boldsymbol{\beta}} := \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{\beta}_j$$

$$\boldsymbol{V}' = \left( \boldsymbol{V}_0^{-1} + J \, \Sigma^{-1} \right)^{-1}$$

$$\Sigma \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\mu} \sim \mathcal{W}^{-1}(\eta', \Psi')$$

$$\eta' = \eta_0 + J$$

$$\Psi' = \Psi_0 + \sum_{j=1}^{J} ( \boldsymbol{\beta}_j - \boldsymbol{\mu})( \boldsymbol{\beta}_j - \boldsymbol{\mu})^T$$

# Recall: Multivariate normal updates (for comparison)

$$\boldsymbol{\mu} \mid \Sigma, \boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{m}', \boldsymbol{V}') \tag{3}$$

$$\boldsymbol{m}' = \boldsymbol{V}' \left( \boldsymbol{V}_0^{-1} \boldsymbol{m}_0 + N \, \Sigma^{-1} \, \bar{\boldsymbol{x}} \right)$$

$$\boldsymbol{V}' = \left( \boldsymbol{V}_0^{-1} + N \, \Sigma^{-1} \right)^{-1}$$

and

$$\Sigma \mid \boldsymbol{\mu}, \boldsymbol{x} \sim \mathcal{W}^{-1}(\nu', \Psi')$$

$$\nu' = \nu_0 + N$$

$$\Psi' = \Psi_0 + \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

# Inference on group-specific regression weights

Key Insight: This part of the model behaves like a Bayesian linear regression model, but where the "prior" mean and variance becomes the mean and variance of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$, the regression weights across groups.

$$\boldsymbol{\beta}_j \mid \Sigma, \boldsymbol{\mu}, \sigma^2, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}'_j, \Sigma'_j)$$

$$\Sigma'_j = \left( \underbrace{\Sigma^{-1}}_{\text{between-group precision}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}_j^T \boldsymbol{X}_j}_{\text{within-group precision}} \right)^{-1}$$

$$\boldsymbol{\mu}'_j = \Sigma'_j \left( \underbrace{\Sigma^{-1} \boldsymbol{\mu}}_{\text{between-group precision-weighted mean}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}_j^T \boldsymbol{y}_j}_{\text{within-group precision-weighted mean}} \right)$$

# Recall: Bayesian linear regression (for comparison)

In standard BLR, the updates on the regression weights are given by

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

where

$$\Sigma = \Bigg( \underbrace{\Sigma_0^{-1}}_{\text{prior precision}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X}}_{\text{data precision}} \Bigg)^{-1}$$

$$\boldsymbol{\mu} = \Sigma \Bigg( \underbrace{\Sigma_0^{-1} \boldsymbol{\mu}_0}_{\text{prior precision-weighted mean}} + \underbrace{\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y}}_{\text{data precision-weighted mean}} \Bigg)$$

(4)

# Inference on observation noise

Key Insight: We estimate the observation noise by simply pooling the residuals across all the groups.

$$\sigma^2 \mid \beta_1, ..., \beta_J, \boldsymbol{y} \sim \mathcal{IG}\left(\frac{1}{2}(\nu_0 + N), \frac{1}{2}\left(\nu_0\sigma_0^2 + \text{SSR}(\beta)\right)\right)$$

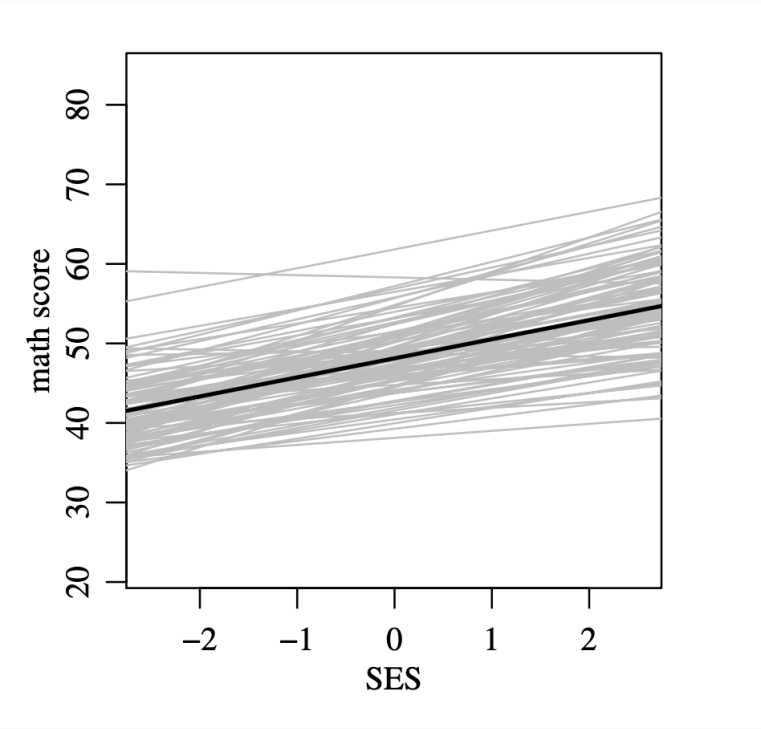
$$N := \sum_{j=1}^{J} n_j$$

$$\text{SSR}(\beta) := \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \beta_j^T \boldsymbol{x}_{ij})^2$$

# Results

The hierarchical model is able to share information across groups.

Posterior expectations of the 100 school-specific regression lines, with the average line given in black.

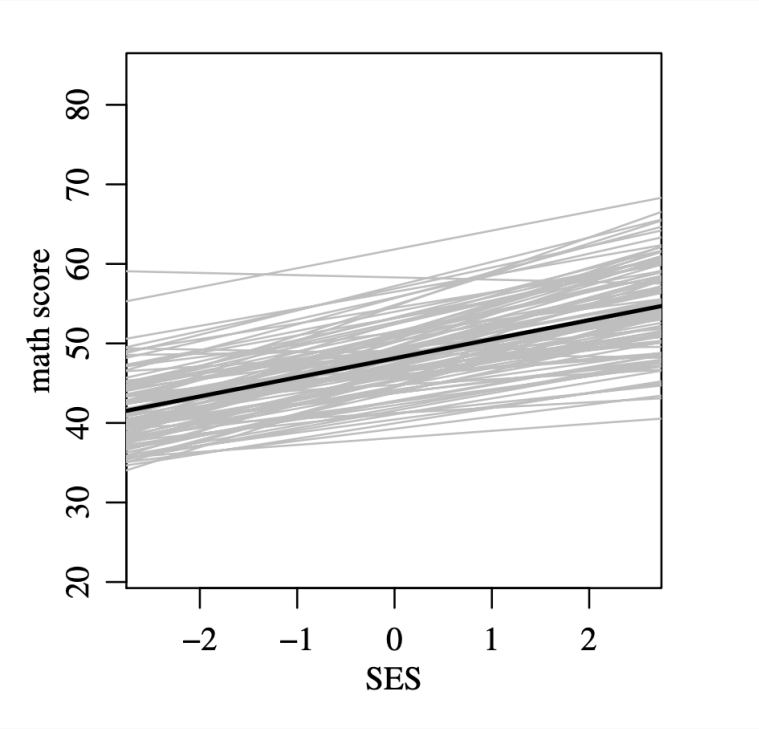



*Observations?*

# Results

The hierarchical model is able to share information across groups.

Posterior expectations of the 100 school-specific regression lines, with the average line given in black.




*Observations?*

- Extreme regression lines are shrunk towards the across-group average. (In particular, hardly any of the slopes are negative after sharing information.)
- Schools with less information have greater shrinkage.