

2023

NYS Graduation Rates



Using Bayesian Inference
to Analyze New York State
Graduation Rates



@asoylatte03



● Rationale

2023

Graduation Rate As A Diagnostic Measure

New York contains the largest public school district in the U.S (NYC) serving over 1 million students across more than 1800 schools.

Graduation Rates allow district and school leadership to be held accountable for student performance and outcomes



2023

Stakeholders

TEACHERS



NONPROFIT ORGS



DISTRICT LEADERS



bayesian inference - bayesian inference - bayesian inference - bayesian inference - bayesian inference

● Data Understanding

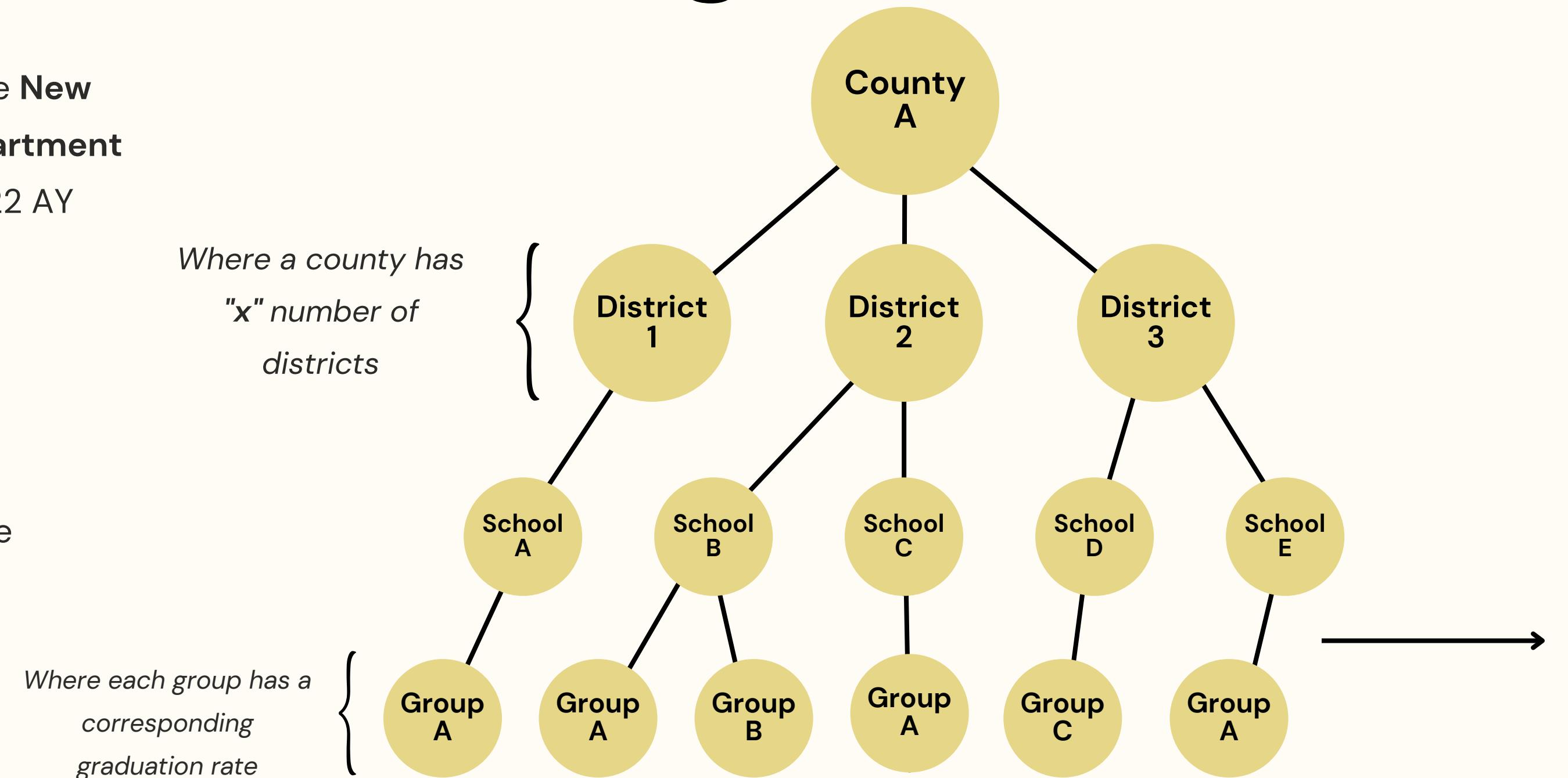
2023

Data was collected from the **New York State Education Department** (NYSED) from the 2021-2022 AY

After feature engineering:

Dataset includes...

- county
- district name
- NRC (Need-to-Resource Capacity Index)
- school name
- student subgroup
- **graduation rate**



*NYSED defines a "graduate" to be anyone who earned a **local** or **Regents diploma** in a given academic year.





2023

Linear Mixed Effects (LME) Model

MODEL #1

Pooled Bayesian Model

MODEL #2

Unpooled Bayesian Model

MODEL #3

Hierarchical Bayesian Model

MODEL #4

Modeling

How do we take advantage of the hierarchical structure of the data to create robust analysis and groupwise comparisons?



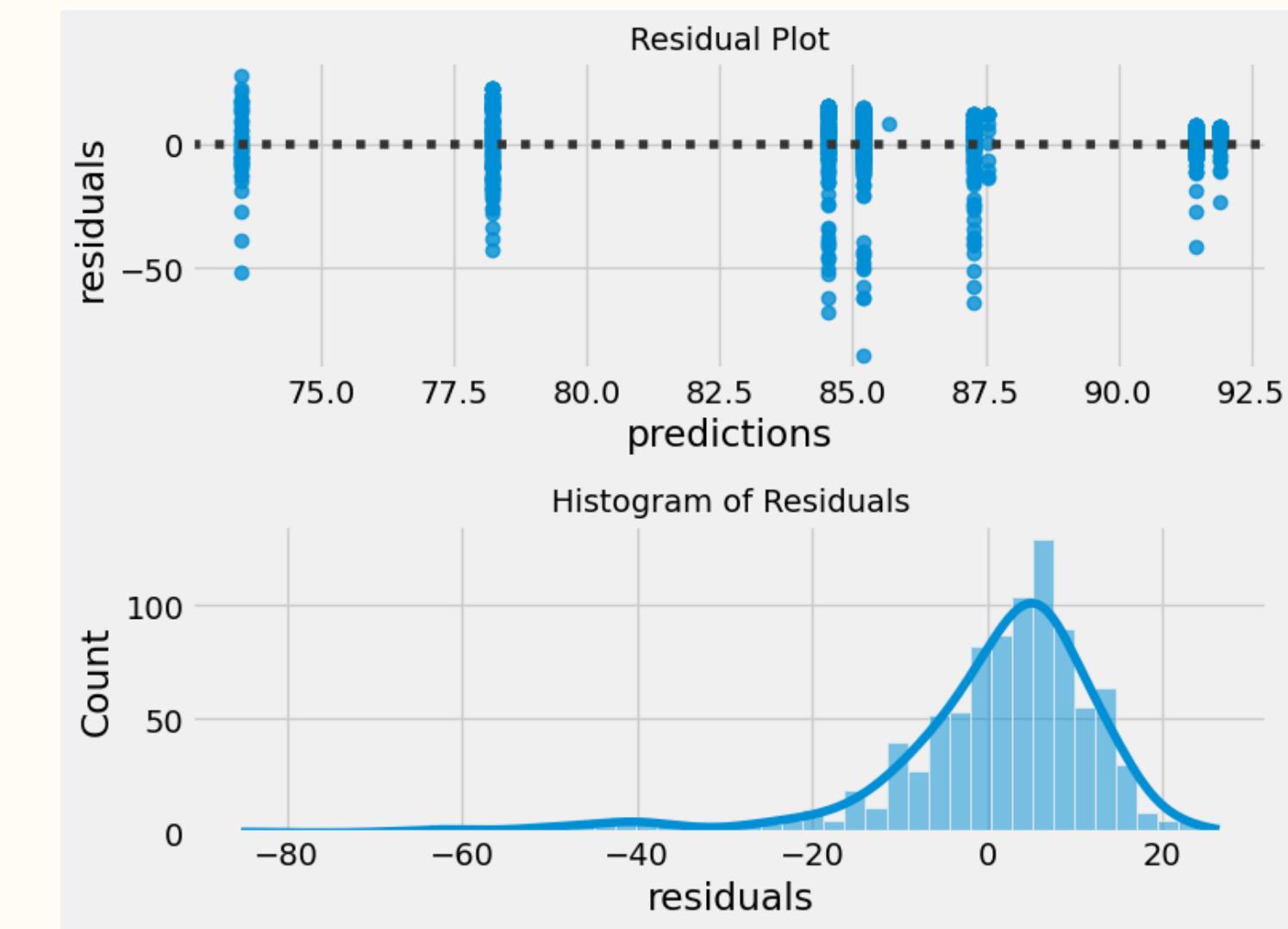
• Linear Mixed Effects (LMER)

Linear Mixed Effect Regression (LMER) is a useful method for dealing with nested data and understanding the significance of **random effects**

After modeling

With an $\alpha = .05$ and accounting for the variance between schools:

- English Language Learners have ~15% lower graduation rate compared to other subgroups
- Hispanic or Latinx students have ~4% lower graduation rate compared to other students
- Students with Disabilities have ~10% lower graduation rate compared to other students



base intercept = ~88%
avg. subgroups/school = 3
mean absolute error ~ 9%





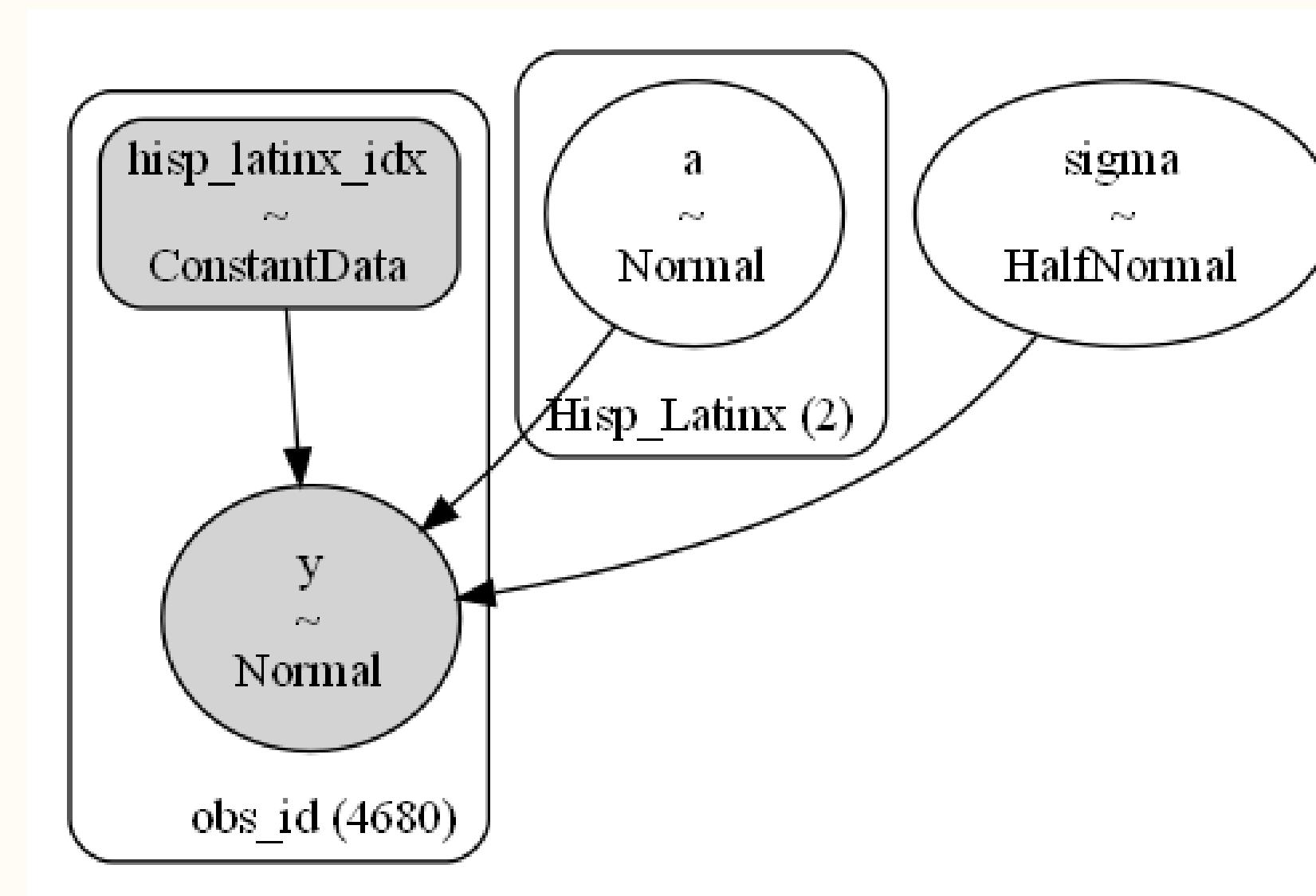
2023

Hierarchical Models !

Bayesian Hierarchical Models allow us to effectively analyze **hierarchical/nested data** where observations are **clustered** within **higher-level** units. They allow us to incorporate prior information and model uncertainty about our parameters

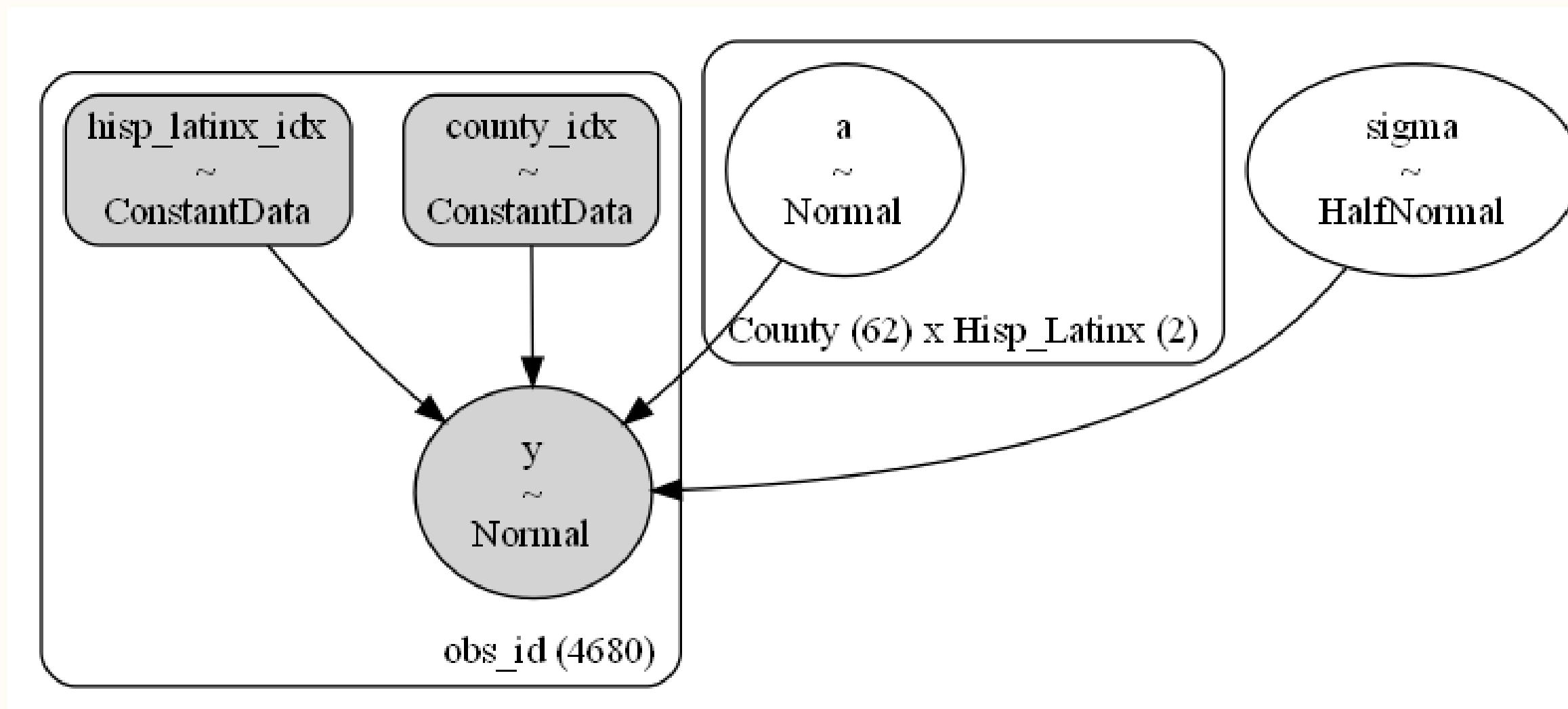


Complete Pooling



The complete pooling model estimates a single graduation rate while treating all counties the same

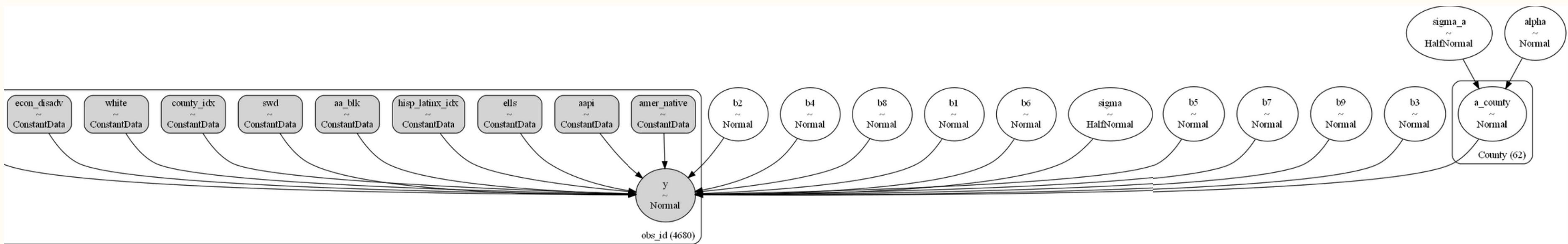
No Pooling/Unpooled



The unpooled approach estimates graduation rates for each county ($n=62$)

This can lead to uncertain estimates where counties may not have adequate measurements

Hierarchical/Multilevel Model

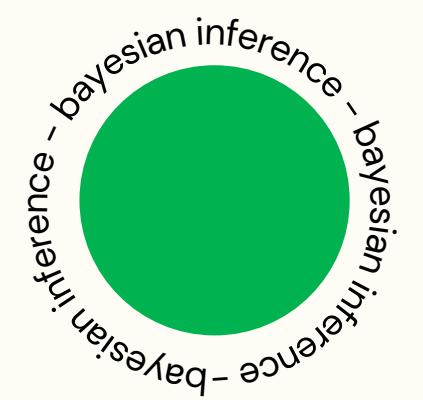
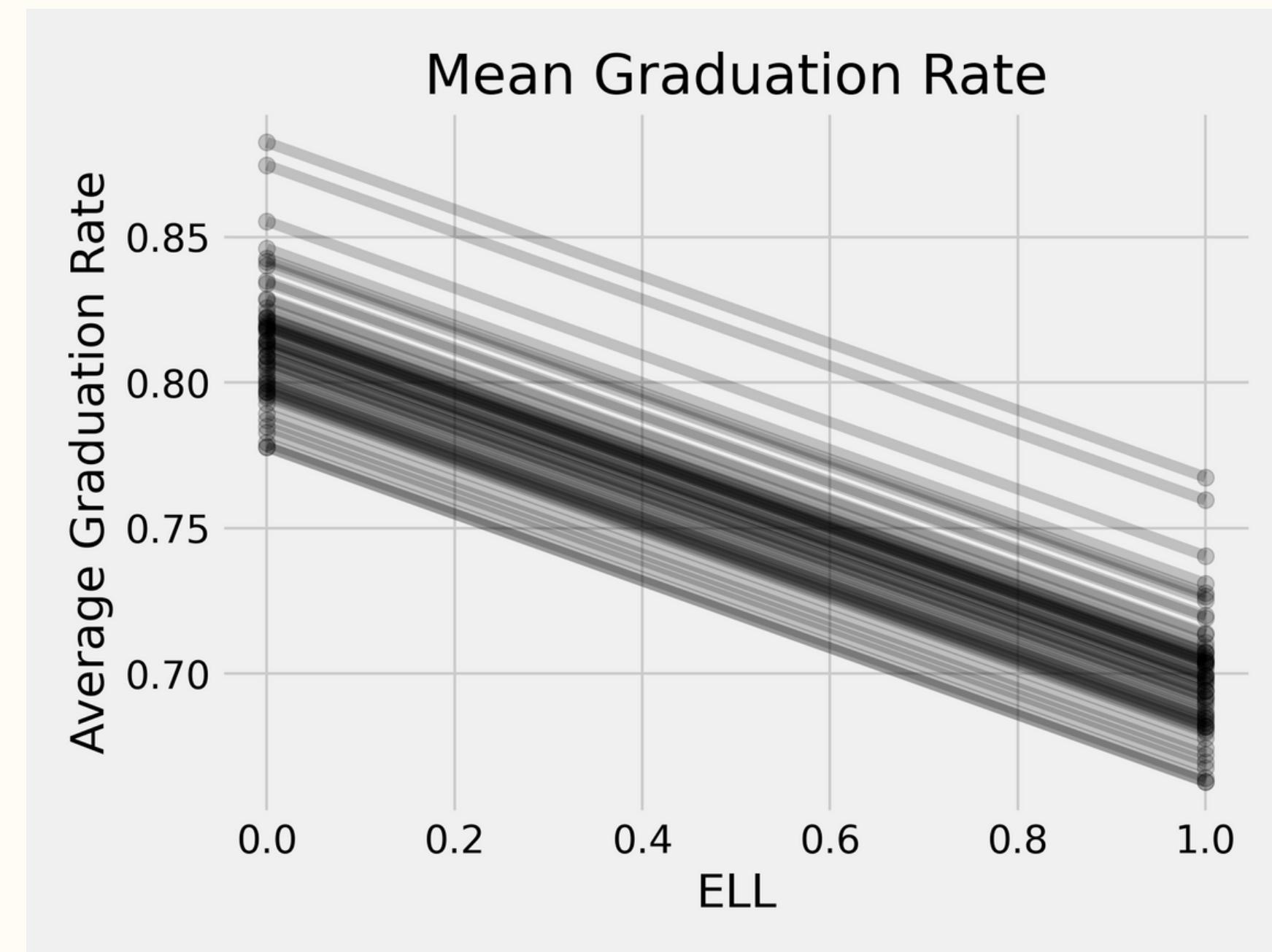


The hierarchical model provides a **balance**
between unpooleed and pooled approaches.

● Hierarchical/Multilevel Model

2023

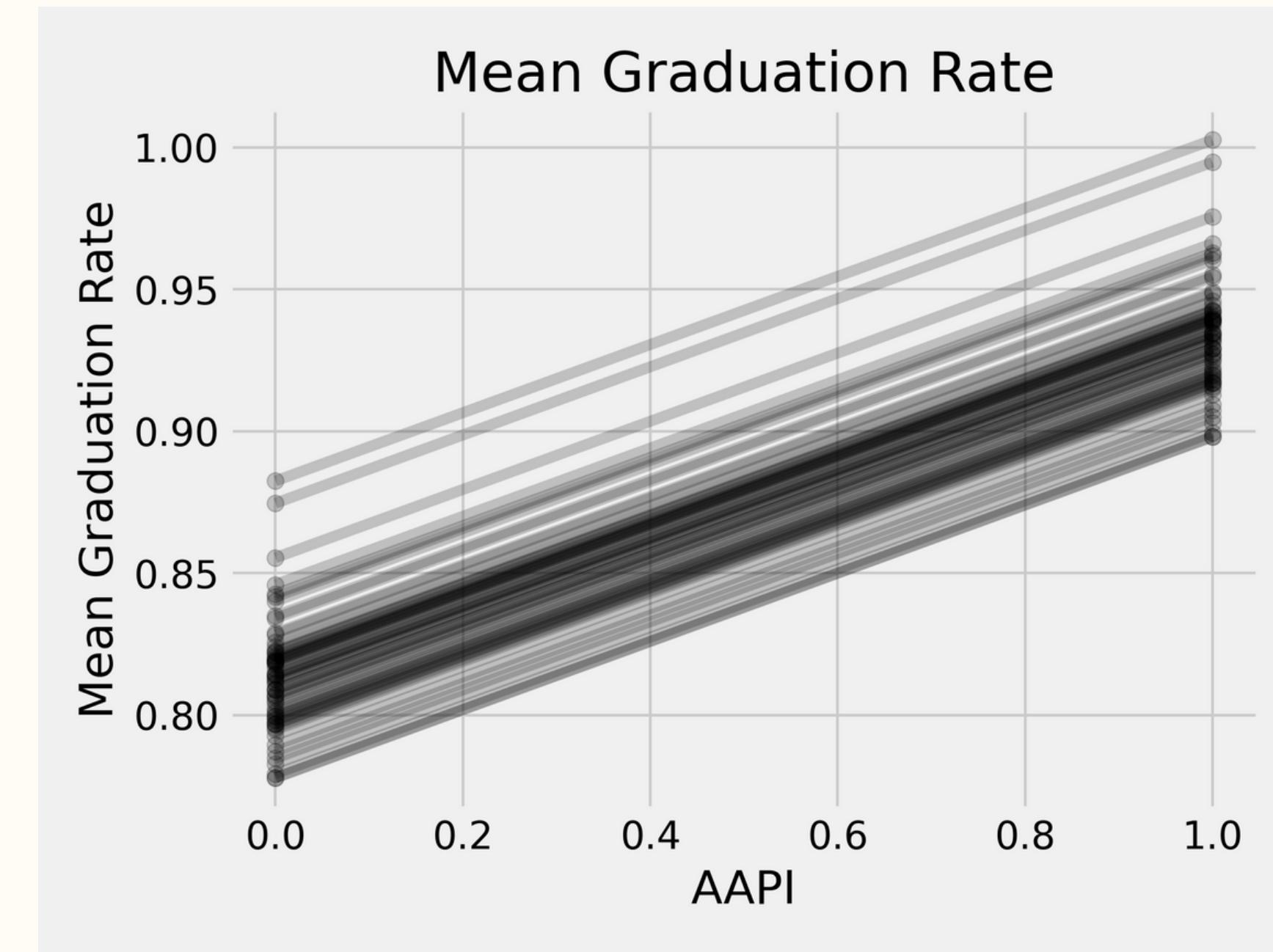
After accounting for the variation across counties, the expected graduation rates for English Language Learners (ELLs) is lower (**68%**) compared to other subgroups (**82%**)



● Hierarchical/Multilevel Model

2023

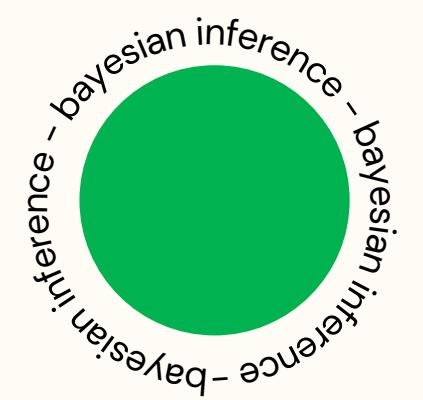
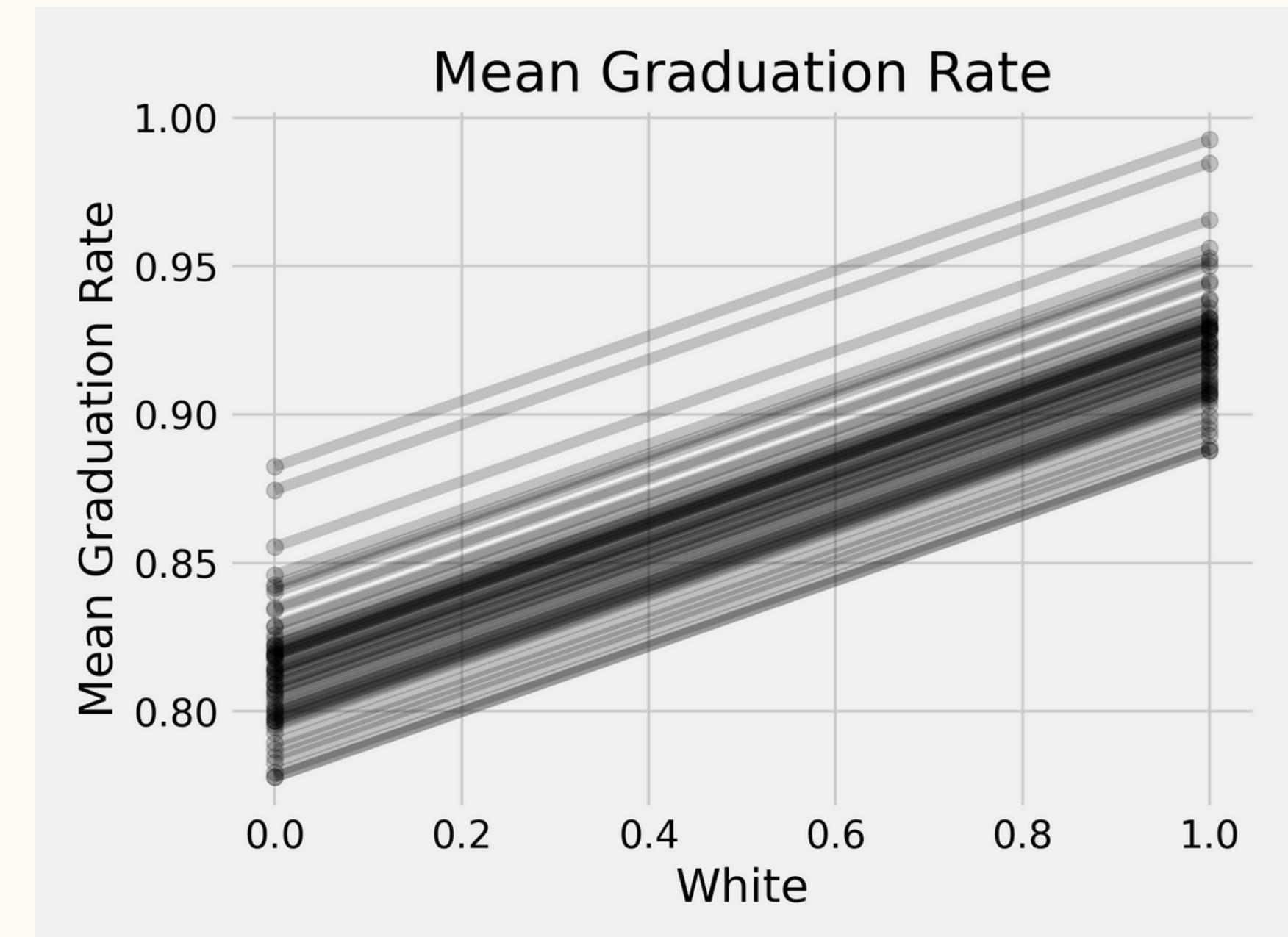
After accounting for the variation across counties, the expected graduation rates for Asian American/Pacific Islander students is higher (**93%**) compared to other subgroups (82%)



● Hierarchical/Multilevel Model

2023

After accounting for the variation across counties, the expected graduation rates for White students is higher (**92%**) compared to other subgroups (**82%**)



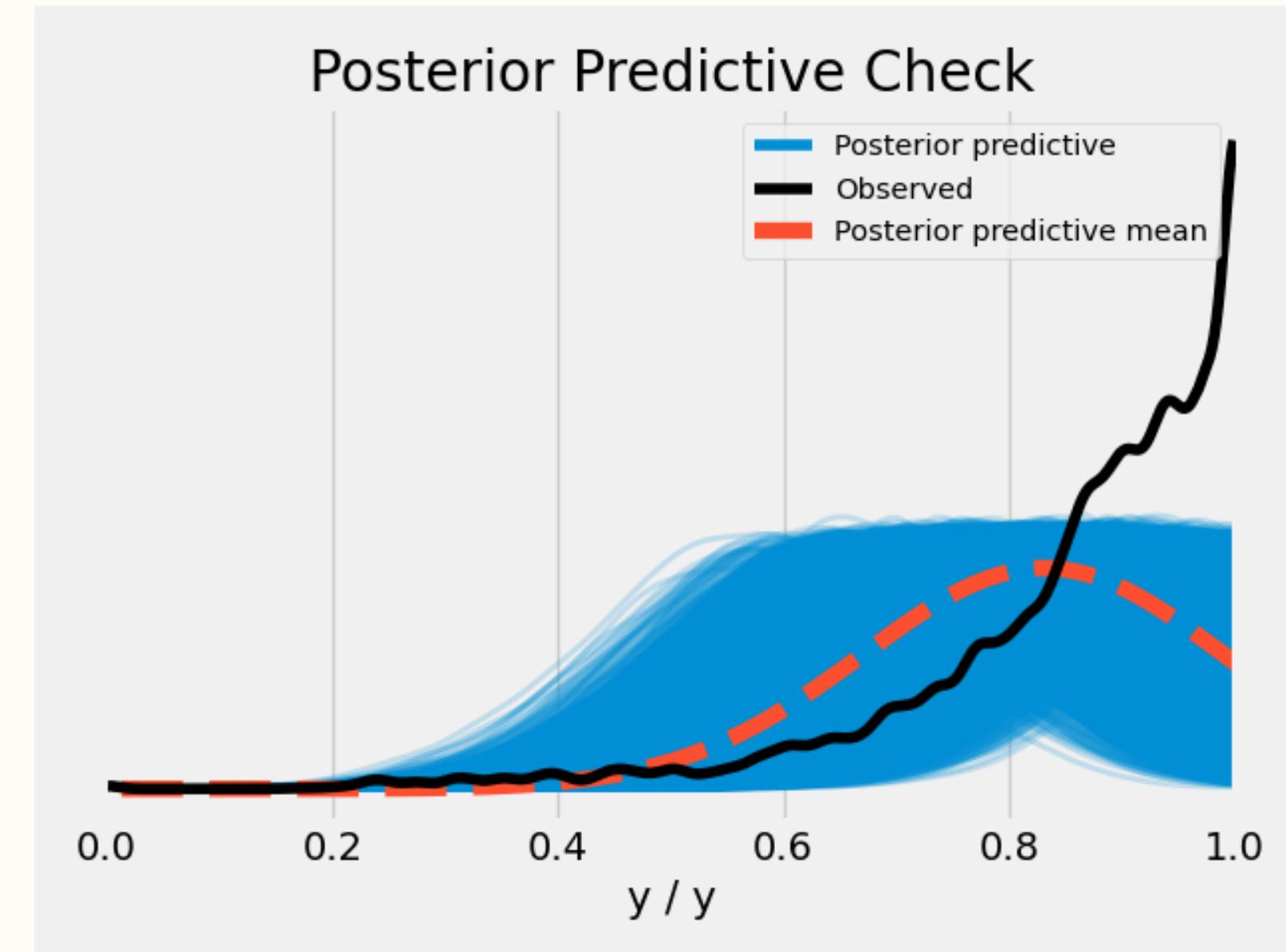
● Model Evaluation – PPC

2023

The Posterior Predictive Check

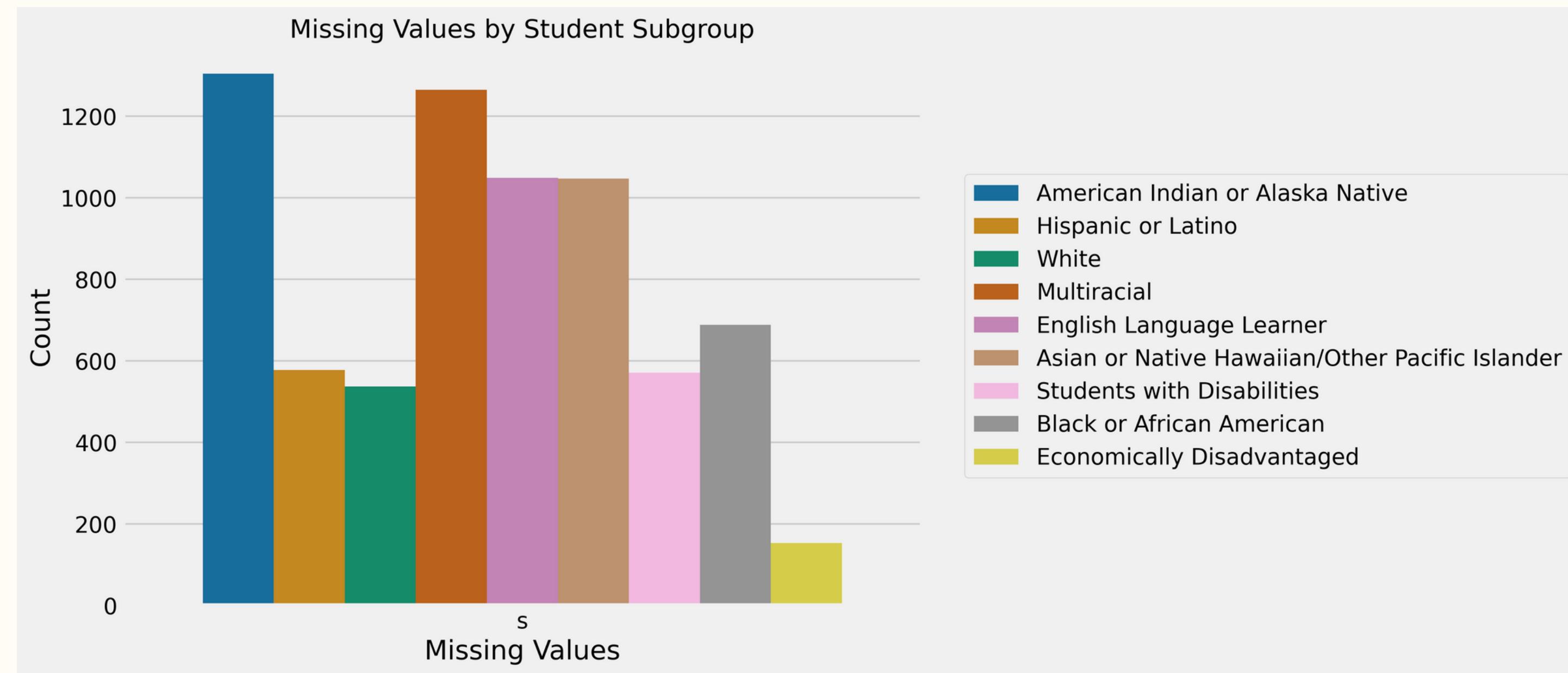
(PPC) provides a visual representation of how a model fits to the original distribution

After **Markov-Chain Monte-Carlo sampling (MCMC)**, the model fails to replicate data right of the tail
(observations where graduation rate > ~83%)



• Data Suppression

2023

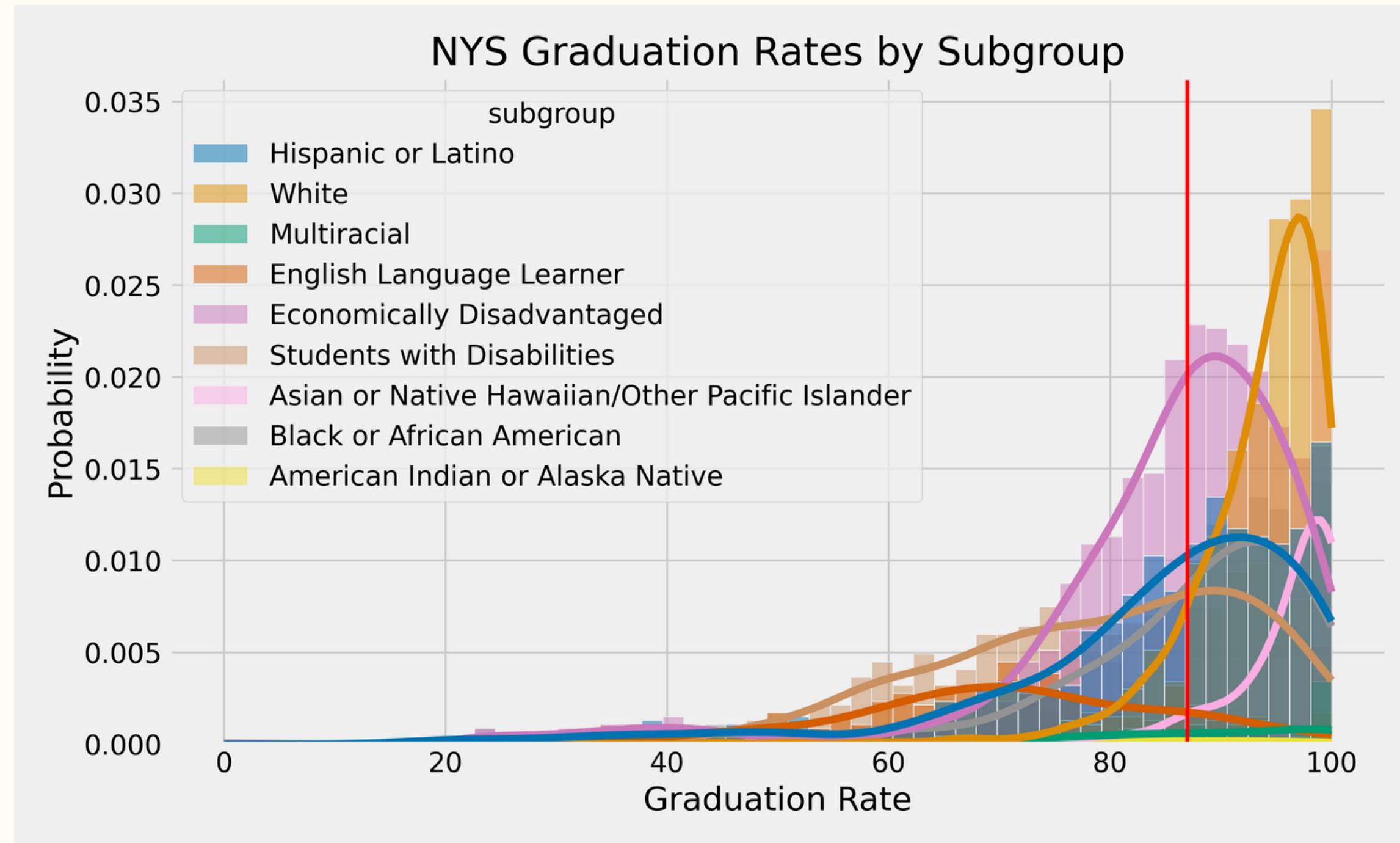


NYSED in compliance with **The Family Educational Rights and Privacy Act (FERPA)** and **NYS** laws suppresses data on subgroups where N<30 to maintain student confidentiality.



● Model Evaluation - Potential Bias

2023



Graduation Rates have a negatively skewed distribution and vary by subgroup. Because of suppressed data, observed data can potentially overrepresent some subgroups compared to others where information is limited/redacted.



• Key Recommendations

Bayesian Inference Results:

- **Inference ≠ Causality**
- Based on data available, we can conclude relative and absolute effects across counties.
 - However, b/c of **bad PPC results**, data might **not** be reflective of the underlying distribution of observed values

Linear Mixed Effects Model:

- **English Language Learners, Hispanic & Latinx, Students with Disabilities** identified as target subgroups

Data Disaggregation

- Breaking up subgroups into much more nuanced groups (e.g., English Language Learners by status & home language spoken)



● Next Steps !

2023

1. Modifying the Hierarchical Structure of the Model:

- a. Instead of a common slope → test varied slope + intercept

2. Adjust Priors & Likelihood Distributions

- a. Leverage expert-knowledge & informative research as part of the model-construction process
- b. Reevaluate prior distributions and likelihood functions (gamma + exponential)

3. Model Complexity

- a. Adding add'tl group-level predictors (at the county-level) (e.g., median household income, poverty rates)



2023



Thank you!

