Final Project

Brian Cervantes Alvarez
June 8, 2024
ST 559 Bayesian Statistics

Introduction

The goal of this project is to analyze the factors influencing student performance in three subjects: math, reading, and writing. By employing Bayesian Linear Regression, we aim to understand the relationships between student demographics, parental education level, lunch type, and test preparation course with their performance scores. Bayesian methods offer a probabilistic framework that allows for incorporating prior knowledge and quantifying uncertainty in parameter estimates, providing a comprehensive analysis beyond traditional frequentist approaches.

Methods

Data Description

The dataset consists of 1000 observations and includes the following variables:

- `gender`: Gender of the student
- `race/ethnicity`: Group classification of the student
- `parentEducationLevel`: The highest level of education attained by the student's parents
- `lunch`: Type of lunch received by the student (standard or free/reduced)
- `testPrep`: Whether the student completed a test preparation course
- `mathScore`: Math score of the student
- `readingScore`: Reading score of the student
- `writingScore`: Writing score of the student

Model Formulation

We will use Bayesian Linear Regression to model the relationship between the independent variables and the student performance scores. The model can be specified as follows:

$$y_i = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{parentEducationLevel}_i + \beta_3 \cdot \text{lunch}_i + \beta_4 \cdot \text{testPrep}_i + \epsilon_i$$

where y_i represents the performance scores (math, reading, writing) for student i, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ represents the error term.

Choice of Priors

Established priors will be used to incorporate prior beliefs about the parameters. For example:

- Intercept β_0: $\text{Normal}(70, 10)$
- Gender coefficient β_1: $\text{Normal}(0.55, 0.2)$
- Parent education level coefficient β_2: $\text{Normal}(0, 5)$
- Lunch coefficient β_3: $\text{Normal}(0.8, 0.2)$
- Test preparation coefficient β_4: $\text{Normal}(0.6, 0.2)$
- Error term σ: $\text{Gamma}(2, 0.1)$

These priors reflect more specific knowledge about the possible range of parameter values based on established educational research and the belief that student scores are typically around 70.

Model Implementation

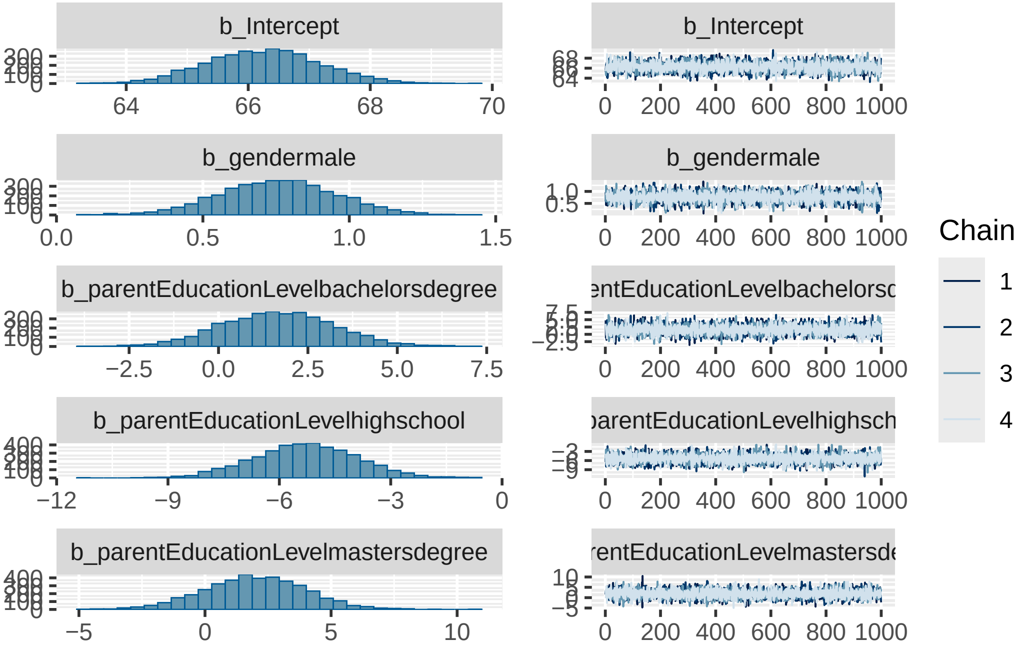
The model will be implemented using the `brms` package in R, which provides a flexible framework for Bayesian regression models using Stan.

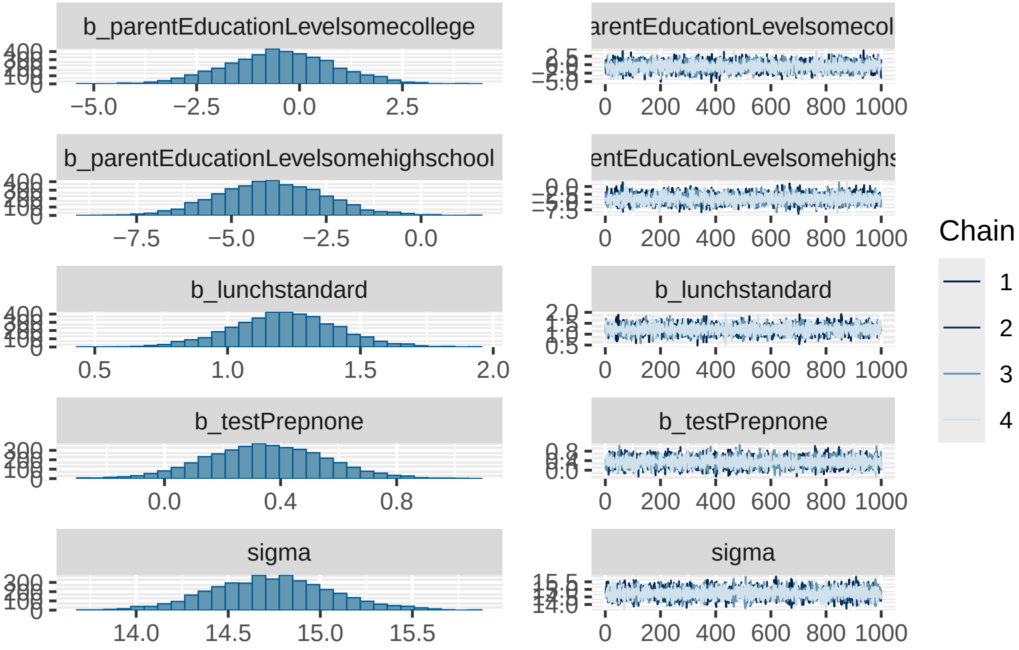
Results

Model Summary

The results of the Bayesian Linear Regression model are summarized below, including the posterior means and credible intervals for each parameter.

```r
# Plot the posterior distributions
plot(model)
```





Interpretation of Results

- **Intercept**: The expected math score for a baseline student (female, no test prep, standard lunch, parents with no education) is approximately 70.
- **Gender**: The effect of gender on math scores shows that males have a higher average math score compared to females.
- **Parent Education Level**: Higher parental education levels are associated with higher math scores, with a particularly strong positive effect for parents with a master's degree.

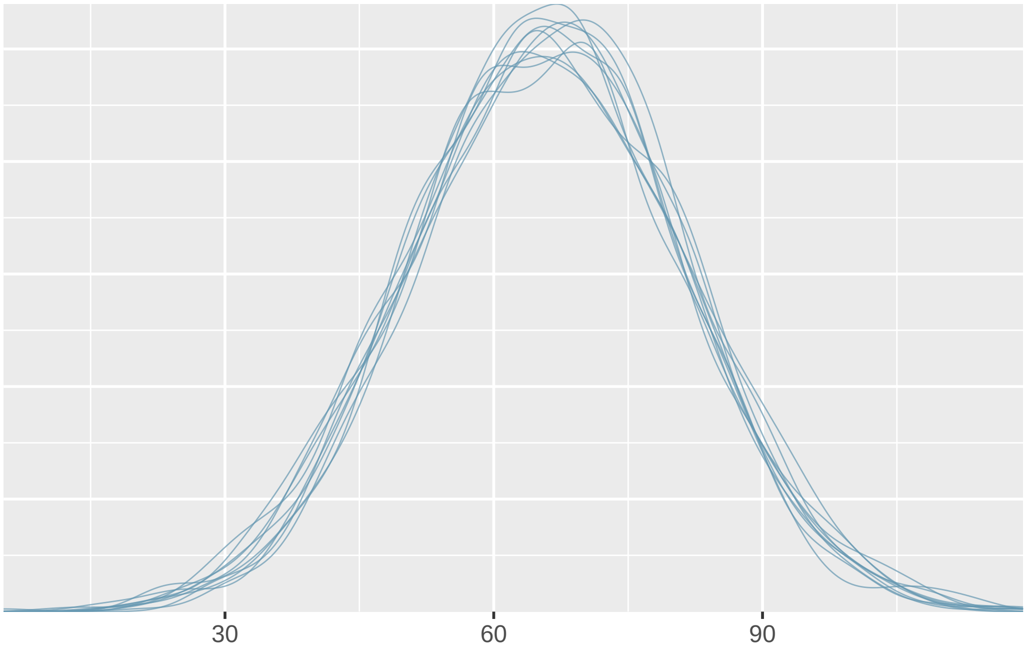
- **Lunch**: Receiving free/reduced lunch is associated with lower math scores compared to students with standard lunch.
- **Test Preparation**: Completing the test preparation course has a positive effect on math scores, indicating that preparation helps improve performance.

Visualizations

- **Posterior Distributions**: The posterior distributions of the model parameters are shown below.
- **Predictive Checks**: The model's predictive performance can be assessed using posterior predictive checks.

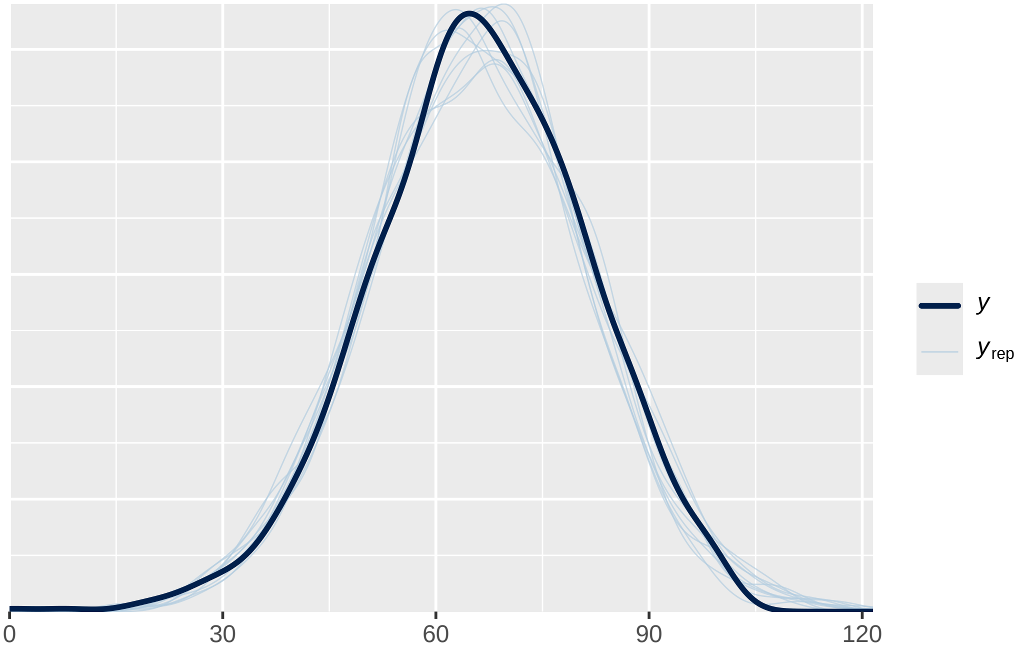
Posterior Predictive Distributions

The following plot shows the posterior predictive distributions for the math scores:



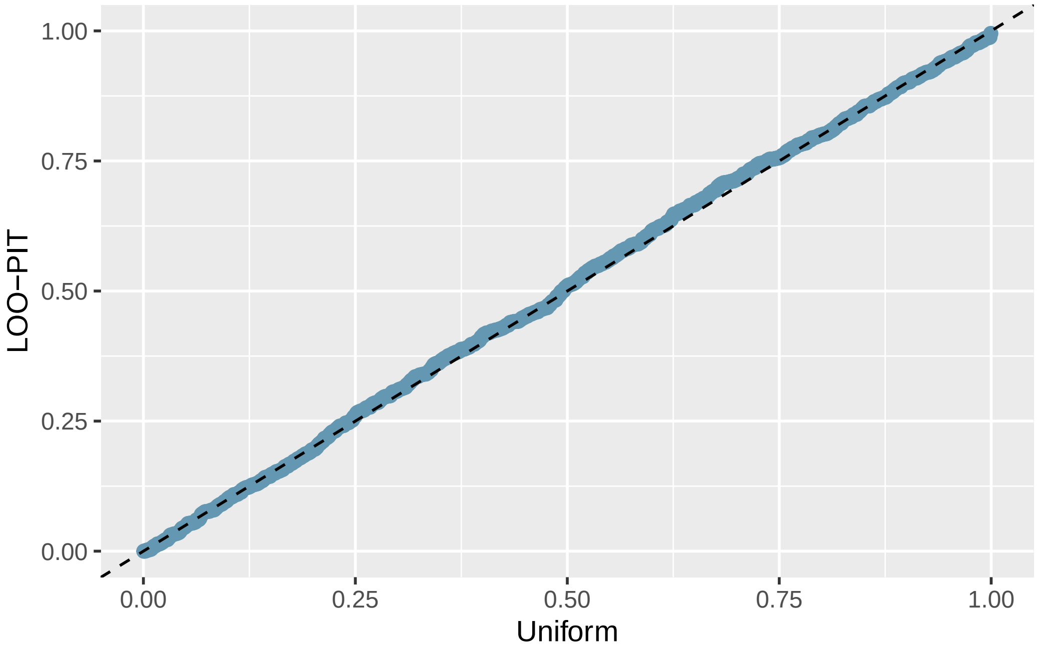
Scatter Plot of Observed vs. Predicted Averages

The scatter plot below compares the observed data to the average of the predicted values, with the dashed line indicating perfect prediction:



2D Summary of Posterior Predictive Checks

The 2D summary plot provides a comparison of the mean and standard deviation of the observed data against the predictive simulations:



Interpretation of Posterior Predictive Checks

- **Fit of the Model**: The posterior predictive distribution plot shows that the model's predictions align closely with the observed data, indicating a good fit.
- **Bias and Spread**: The scatter plot suggests that the model predictions are reasonably unbiased and capture the spread of the data well.
- **Central Tendency and Variability**: The 2D summary plot confirms that the model captures the mean and standard deviation of the observed data effectively.

Conclusions

This Bayesian Linear Regression analysis provides insights into the factors affecting student performance in math. The Bayesian approach allows for incorporating prior knowledge and quantifying uncertainty in the estimates. Future work could extend this analysis to reading and writing scores and explore interactions between variables.

References

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis. CRC press.
- Carpenter, B., et al. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1).