Bayesian Analysis of Math Scores

Brian Cervantes Alvarez
June 9, 2024
ST 559 Bayesian Statistics

Introduction

The goal of this project is to analyze the factors influencing student performance in math. Specifically, the research question we aim to address is: "How do student demographics, parental education level, lunch type, and test preparation course influence student performance scores in math?" By employing Bayesian Linear Regression, we seek to understand the relationships between these variables and performance scores. In essence, 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

The dataset used in this analysis, obtained from Kaggle, consists of 1,000 observations of student performance across several demographic and educational variables. These variables include gender, race/ethnicity, the highest level of education attained by the student's parents, the type of lunch received by the student (standard or free/reduced), and whether the student completed a test preparation course. The performance scores in math, reading, and writing were the dependent variables, allowing us to investigate how these factors might influence student outcomes in these areas.

The linear regression model can be expressed as:

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

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

The priors chosen for this model reflect established knowledge from educational research. For instance, the intercept, representing the expected score for a baseline student, was given a Normal(70, 10) prior, reflecting the belief that average scores are around 70. Coefficients for other variables, such as gender, parental education, lunch type, and test preparation, were also assigned normal

priors based on the assumption of their influence on performance. The error term σ was given a Gamma(2, 0.1) prior.

The model was implemented using the `brms` package in R, which provides a flexible framework for specifying and fitting Bayesian regression models using Stan. After preprocessing the dataset and factorizing the categorical variables, we implemented the chosen priors. The model fitting process involved running Markov Chain Monte Carlo simulations to obtain the posterior distributions of the model parameters.

Results

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

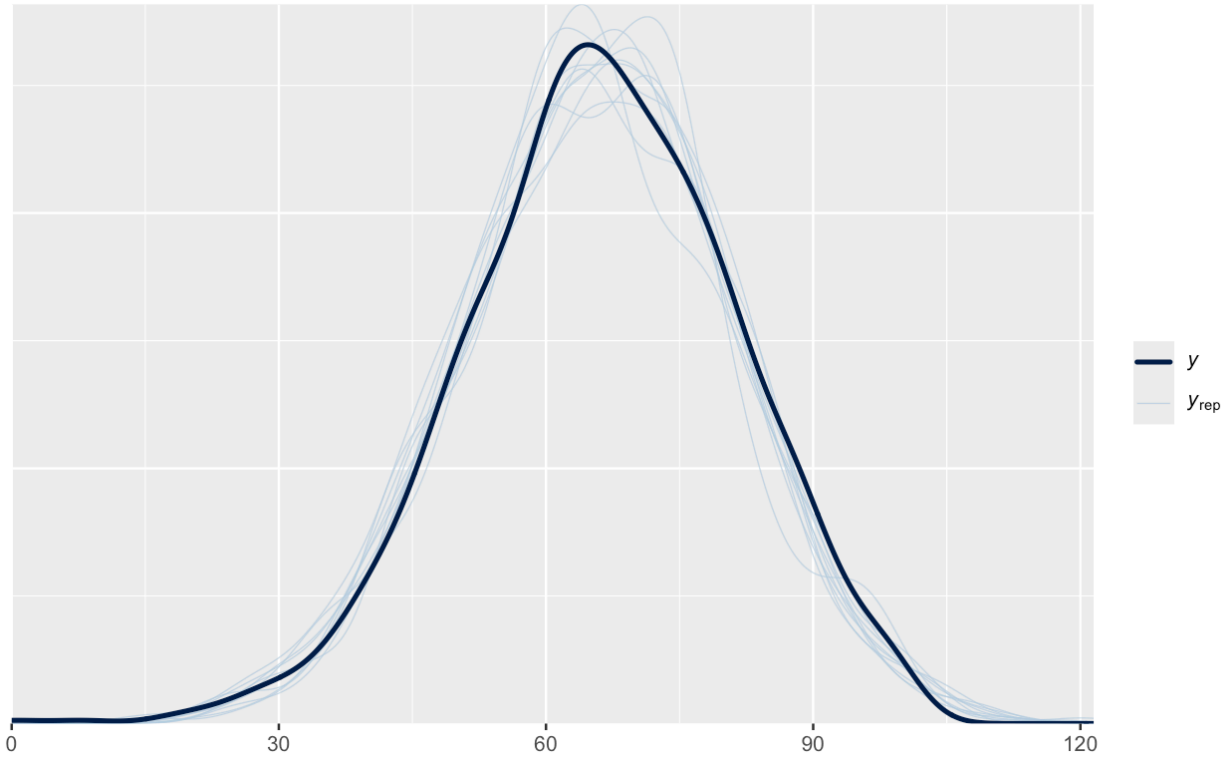

Figure 1: Posterior Predictive Distributions

This plot illustrates the model's ability to predict the observed distribution of math scores. The alignment of the predictive distributions (light blue lines) with the observed data (dark blue line) indicates that the model provides a good fit.

Next, the results of the Bayesian Linear Regression model provided insights into the effects of different variables on math scores. The table below summarizes the posterior means and credible intervals for each parameter, indicating the strength and direction of their influence.

Parameter	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	66.28	0.92	64.48	68.13	1.00	4782	3206
gendermale	0.76	0.20	0.38	1.15	1.00	7884	2943
parentBachelors	1.66	1.54	-1.35	4.64	1.00	5756	3546
parentHS	-5.33	1.37	-7.96	-2.67	1.00	4970	3422
parentMasters	1.95	1.90	-1.84	5.67	1.00	5360	2914
parentSomeColl	-0.50	1.31	-2.96	2.14	1.00	5553	3408
parentSomeHS	-3.95	1.34	-6.66	-1.36	1.00	5477	3705
lunchstandard	1.22	0.20	0.83	1.60	1.00	6366	3181
testPrepnone	0.35	0.20	-0.03	0.74	1.00	8299	3013
sigma	14.74	0.33	14.13	15.40	1.00	7759	3209

The intercept indicates that the expected math score for a baseline student (female, no test preparation, standard lunch, parents with no education) is approximately 66.28. The gender coefficient shows that male students tend to score 0.76 points higher on average than female students. Parental education level significantly affects math scores, with bachelor's and master's degrees being associated with higher scores, while having only a high school education or some high school education is associated with lower scores. Standard lunch provision is positively associated with math scores, while the completion of test preparation shows a modest positive effect.

Conclusion

This Bayesian Linear Regression analysis reveals that gender, parental education, lunch type, and test preparation significantly influence math scores. Male students score 0.76 points higher than female students on average. Higher parental education levels correlate with better scores, while standard lunch provision and test preparation completion are associated with modestly higher math scores. Future work could extend this analysis to other subjects like reading and writing to see if the same factors influence performance across different areas. Additionally, exploring interactions between variables, such as how the combination of parental education and test preparation might jointly affect student scores, could provide more nuanced insights. However, the findings are specific to the dataset used and may not generalize to all student populations, as factors affecting performance can vary widely. Furthermore, while the analysis identifies associations between variables and math scores, it does not establish causality. Addressing these limitations and exploring further extensions can enhance our understanding of the factors influencing student performance.

References

Kaggle Dataset: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams