Final Project

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 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

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:

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

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

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.

Regression Coefficients:

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
parentEduLvlbachelorsdegree	1.66	1.54	-1.35	4.64	1.00	5756	3546
parentEduLvlhighschool	-5.33	1.37	-7.96	-2.67	1.00	4970	3422
parentEduLvlmastersdegree	1.95	1.90	-1.84	5.67	1.00	5360	2914
parentEduLvlsomecollege	-0.50	1.31	-2.96	2.14	1.00	5553	3408
parentEduLvlsomehighschool	-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

Further Distributional Parameters:

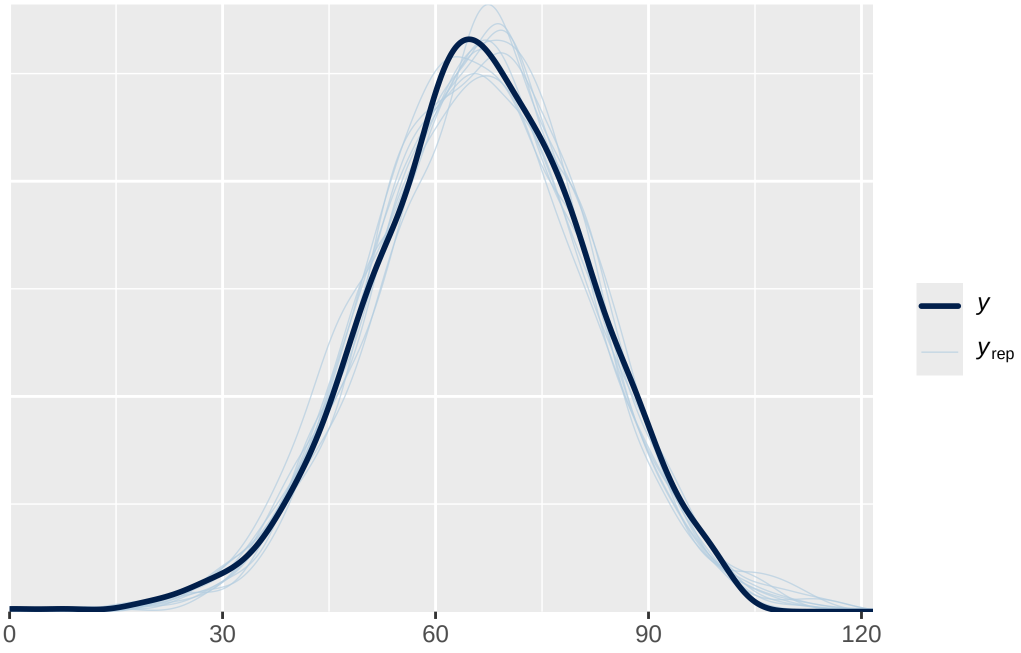
Parameter	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	14.74	0.33	14.13	15.40	1.00	7759	3209

Interpretation of Results

- **Intercept**: The expected math score for a baseline student (female, no test prep, standard lunch, parents with no education) is approximately 66.28.
- **Gender**: The effect of gender on math scores shows that males have a higher average math score by about 0.76 points compared to females.
- **Parent Education Level**: Higher parental education levels generally lead to higher math scores, although the results vary with significant positive effects for bachelor's and master's degrees.
- **Lunch**: Receiving free/reduced lunch is associated with a lower math score compared to students with standard lunch.
- **Test Preparation**: Completing the test preparation course has a positive effect on math scores, although this effect is not strongly significant.

Posterior Predictive Distributions

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



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).