Male students with parents of master degree achieve higher math score*

Yihang Cai

March 19, 2024

1 Introduction

Understanding the factors that influence academic performance has been a critical area of educational research, as it can inform strategies to enhance learning outcomes and identify areas where support is needed. The Section 2 section introduces the dataset obtained from kaggle. The Section 3 section justifies the choice of the model and then summaries the outcome of the model.

2 Data

The dataset hosted on Kaggle, titled "Student Study Performance," provides a rich source of information to analyze various aspects that may affect a student's academic scores. This dataset encompasses a range of variables including demographic details, parental level of education, test preparation courses, and students' math scores.

Data for this study were prepared and analyzed using R (R Core Team 2023), grasping several packages including Tidyverse (Wickham et al. 2019a) for data manipulation, Janitor (Firke 2023) for data cleaning, Readr (Wickham, Hester, and Bryan 2023) for data import, Dplyr (Wickham et al. 2023) for data manipulation, Knitr (Xie 2014) for dynamic reporting, Modelsummary (Arel-Bundock 2022) for summarizing model outputs, and Rstanarm (Goodrich et al. 2022) for Bayesian modeling. Some of the codes are adapted from Telling Stories with Data (Wickham et al. 2019b)

*Code and data are available at: https://github.com/peachvegetable/STA302-mini-essay-10

Table 1: top 10 rows of the dataset

Gender	Level of Parent Education	Test Preparation	Math Score
female	bachelor's degree	none	72
female	some college	completed	69
female	master's degree	none	90
male	associate's degree	none	47
male	some college	none	76
female	associate's degree	none	71
female	some college	completed	88
male	some college	none	40
male	high school	completed	64
female	high school	none	38

Table 1 shows the top 10 rows of the selected data, displaying students' math score, gender, parents' level of education and test preparation status.

3 Model

3.1 Model set-up

Define y_i as the math score. Then β_0 is the intercept and β_1 to β_3 is coefficient to the variables.

$$y_i|\lambda_i, \theta \sim \text{NegativeBinomial}(\mu_i, \theta)$$
$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{parental\_level\_of\_education}_i + \beta_3 \times \text{test\_preparation\_course}_i$$
$$\beta_0 \sim \text{Normal}(0, 2.5)$$
$$\beta_1 \sim \text{Normal}(0, 2.5)$$
$$\beta_2 \sim \text{Normal}(0, 2.5)$$
$$\beta_3 \sim \text{Normal}(0, 2.5)$$

3.2 Model summary

Table 2 displays the summary of the model. Intercept: The model's intercept, 4.229, is the log count of the expected math_score when all other predictors are held at their reference levels (which typically means 'female' for gender, 'none' or a baseline category for parental level of education, and 'completed' for test preparation course).

Table 2: Negative binomial model summary

	Negative binomial
(Intercept)	4.229
	(0.020)
gendermale	0.082
	(0.015)
parental_level_of_educationbachelor's degree	0.021
	(0.026)
parental_level_of_educationhigh school	−0.085
	(0.023)
parental_level_of_educationmaster's degree	0.037
	(0.035)
parental_level_of_educationsome college	−0.010
	(0.023)
parental_level_of_educationsome high school	−0.074
	(0.024)
test_preparation_coursenone	−0.082
	(0.016)
Num.Obs.	1000
Log.Lik.	−4148.533
ELPD	−4156.0
ELPD s.e.	35.6
LOOIC	8312.0
LOOIC s.e.	71.2
WAIC	8312.0
RMSE	14.44

Table 3: Mean and variance of students' math score

Mean	Variance
66.1	229.9

gendermale: The coefficient for 'gendermale' is 0.082. This suggests that, holding all else constant, being male is associated with a log count increase in the math_score of about 0.082 as compared to being female.

parental_level_of_education: This set of coefficients compares different levels of parental education to the reference category. For example:

bachelor's degree: The coefficient is 0.021, meaning that having a parent with a bachelor's degree is associated with a log count increase in the math_score of about 0.021 as compared to the reference category. high school: The coefficient is -0.085, meaning that having a parent with only a high school education is associated with a log count decrease in the math_score of about 0.085 as compared to the reference category. The other education levels can be interpreted similarly. A positive coefficient indicates an increase in log count score compared to the reference, and a negative coefficient indicates a decrease. test_preparation_course: The coefficient for 'test_preparation_coursenone' is -0.082. This indicates that not taking a test preparation course is associated with a log count decrease in the math_score of about 0.082 as compared to those who did complete a test preparation course.

3.3 Model justification

Table 3 shows the mean and variance of students' math scores can be used to justify the choice of a negative binomial regression model over Poisson or logistic regression models when dealing with count data that exhibit overdispersion. The Poisson regression model assumes that the mean and variance of the count data are equal (equidispersion). If we were to use a Poisson model for data where the variance significantly exceeds the mean, as it does here (mean = 66.1, variance = 229.9), it would likely underestimate the variance. This underestimation can lead to confidence intervals that are too narrow and p-values that are too small, increasing the risk of Type I errors (incorrectly rejecting the null hypothesis). The negative binomial regression model is a generalization of the Poisson regression model that includes an additional parameter to model the overdispersion. This extra parameter allows the variance to be greater than the mean. Given that the variance in your data is much greater than the mean, the negative binomial model is a suitable choice because it can accommodate this overdispersion, providing more accurate standard errors and confidence intervals than a Poisson model. Logistic regression is used for binary outcome variables, not for count data. Since the outcome variable in question is a count (math score), logistic regression is not an appropriate model for this data. The choice between a logistic model and a model for count data would typically depend on the nature of the dependent variable. Since we are dealing with count data rather than binary outcomes, logistic regression is not applicable.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.