Math Performance in Relation to School and Social Factors

Isabel Gomez, Dhajanae Sylvertooth, Nicholas Hartman, Ritoban Kundu

Github Repository

The Github repository used in this project can be accessed at https://github.com/igomez39/BIOS625.

Work Contributions

Everyone: Report writing, Project planning
Isabel:
Dhajanae:
Nicholas: Data long format code, Gini Index bootstrap analysis, Cluster computing for GLS model
Ritoban:

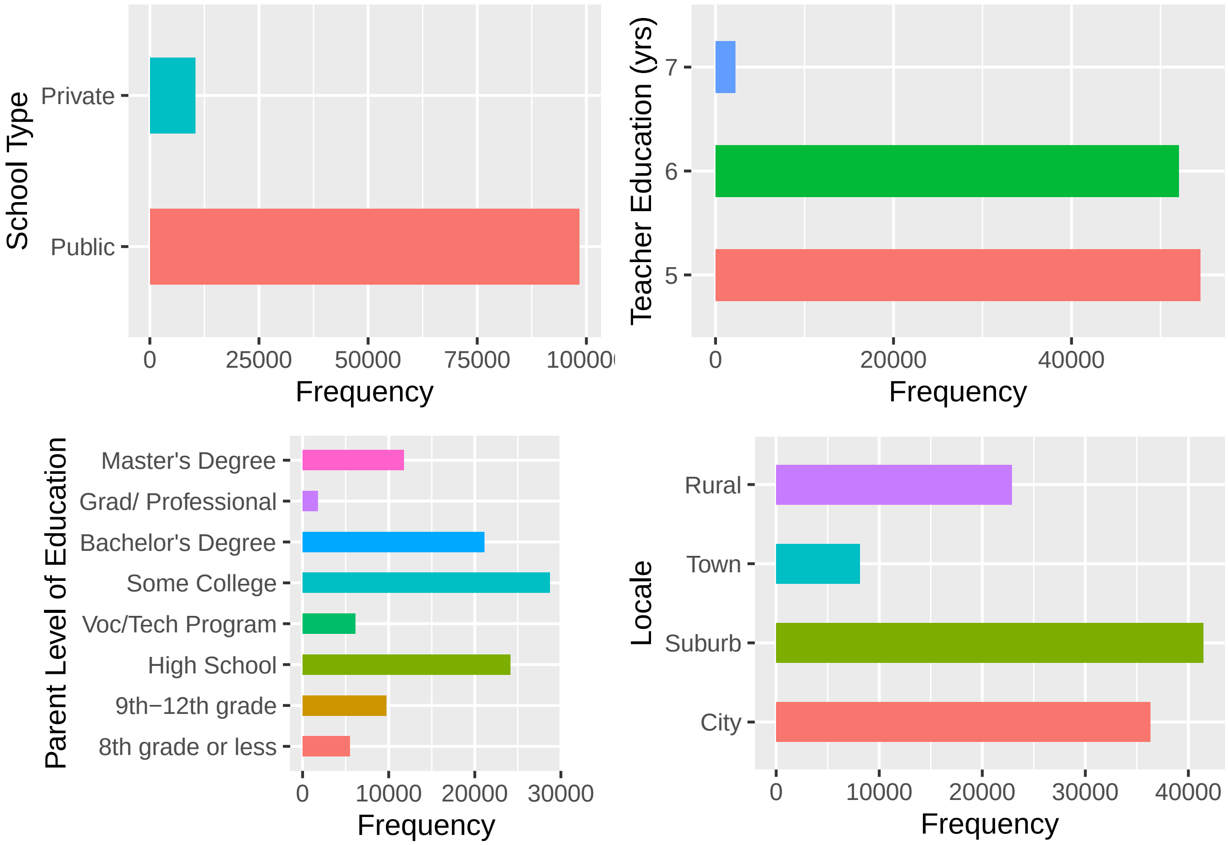
Introduction

Academic proficiency during the formative years of schooling has widespread implications on future development [1]. In this report, we focus on math performance in elementary school students over time, and study the relationships between math proficiency and several social and school characteristics. Using the Gini Index, nonparametric estimation, and longitudinal modeling techniques, we explore several different aspects of math proficiency, such as inequality and progress over time.

The data for this analysis come from the Early Childhood Longitudinal Study (ECLS), which was a prospective cohort study of elementary school students from 2011-2016 [2]. Students were followed from kindergarten to fifth grade, and math performance was assessed each year. In addition, detailed surveys were conducted over time to collect data on a wide range of social factors. The public-use dataset contains information on over 26,000 variables for 970 schools and 18,174 individual students.

Descriptive Plots

We first made descriptive plots to visualize the demographics of the study sample and basic trends in the data. The figures below show frequencies by different categories of interest. Most students in the cohort attended a public school, and both the teachers and parents in this sample tended to be highly educated. However, there was still enough diversity in these variables to study the differences across groups. The cohort included a well-balanced mix of students from urban, suburban, and rural areas, which is a strength of the study design.



Math Score Inequality

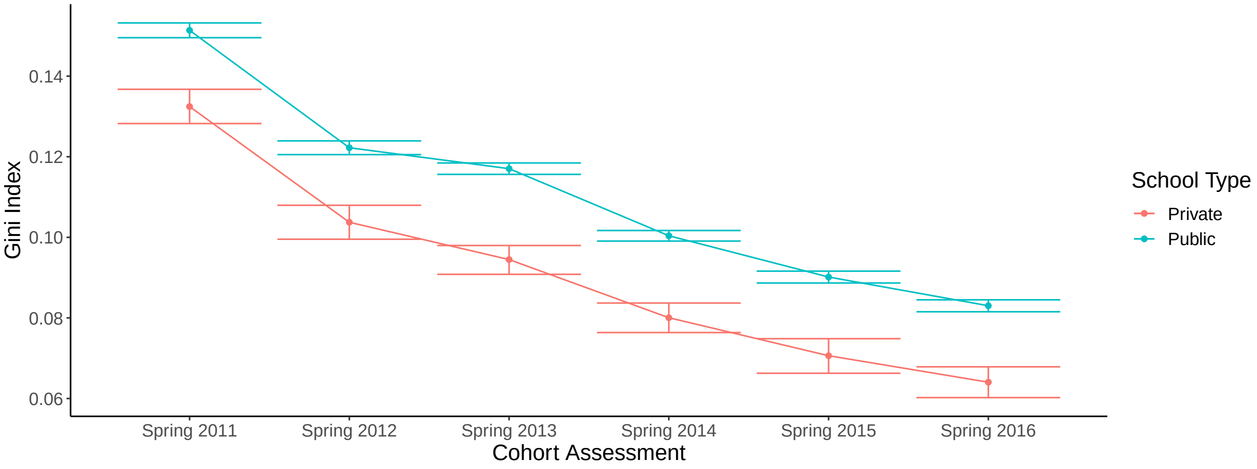
We studied the amount of inequality in math performance and how this changes over time within public and private elementary schools. To accomplish this, we used the Gini Index, which is a well-known measure of inequality and is commonly used to describe wealth inequality. The formula for the Gini Index is [3]:

$$\text{Gini} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}},$$

where x represents the math score. Higher Gini Index values correspond to more inequality in the math scores. For this analysis, we used a re-scaled version of the math scores since x must be positive in the Gini Index formula. There are two main computational challenges that arise in Gini Index estimation [3]:

1. Due to a lack of theoretical formulas for the Gini Index standard error, nonparametric bootstrapping is needed to estimate the uncertainty in the Gini Index.
2. The Gini index involves many pairwise comparisons for large datasets.

We addressed the challenges of (2) by using a vectorized implementation of the Gini Index from the reldist package in R [4]. Then, to further speed up the resampling in (1), we used parallel computing. For each type of school (Public vs. Private) and for each timepoint, we resampled the corresponding data with replacement 10,000 times and computed the Gini Index on each of these bootstrap samples. A 95% confidence interval for each point estimate was computed from the quantiles of the bootstrap distribution. The figure below shows the Gini Index estimates and bootstrap 95% confidence intervals over time, stratified by school type:



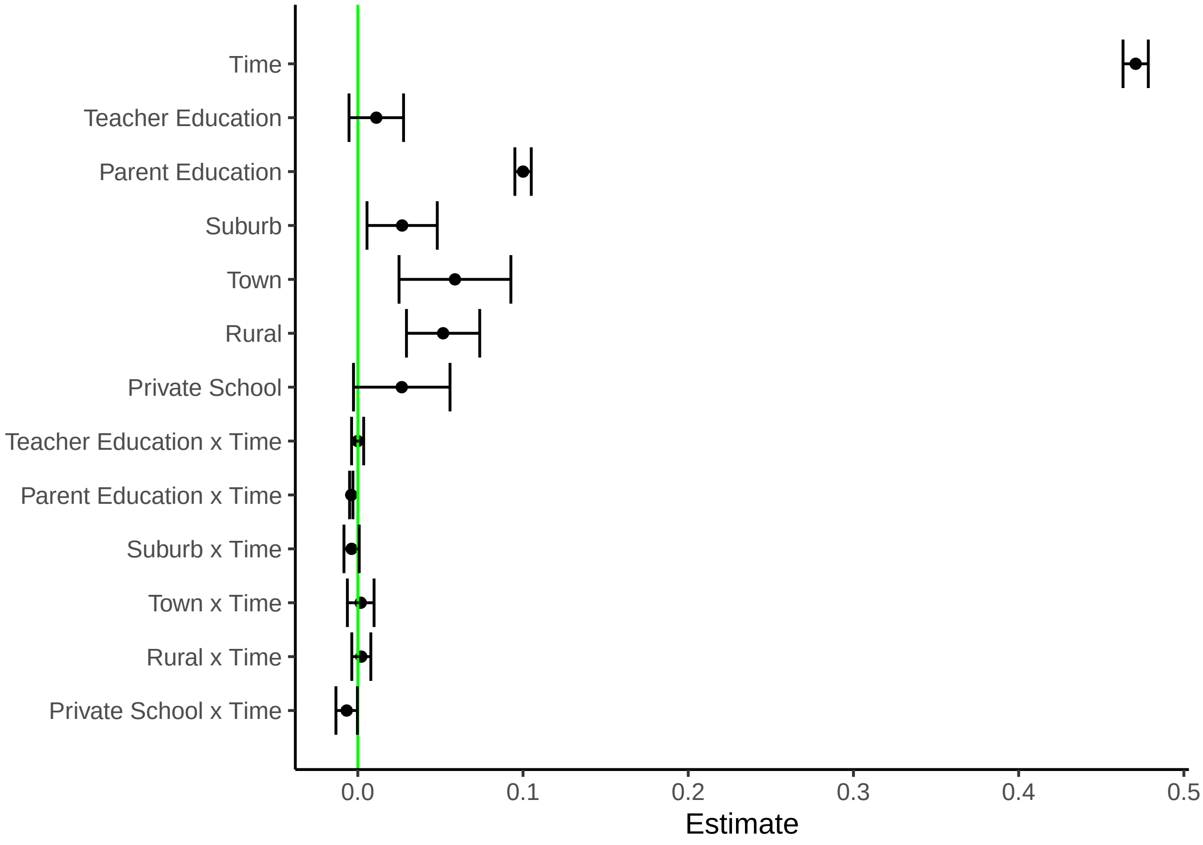
We find that the Gini Index decreases over time within both public and private schools. One interpretation of this result is that the kindergarten students with lower scores were able to catch up to the high-performers by fifth grade. We also observed that there is more inequality in math scores within public schools compared to within private schools. The magnitude of this difference remained fairly constant over time, and since the 95% confidence intervals do not overlap, we can conclude that these differences are statistically significant.

Longitudinal Models

To study the relationships between social factors and math performance over time, we used a Generalized Least Squares (GLS) approach. We chose to use GLS over mixed models because we are mainly interested in the marginal associations. No structure was assumed for the covariance matrix of scores over time because it resulted in a substantially better model fit (and AIC) compared to other covariance structures (such as AR1). We specified the marginal model as:

$$\begin{aligned} E[\text{Score}_{ij}] = \beta_0 &+ \beta_1 \text{Time}_{ij} + \beta_2 I(\text{School}_i = \text{Private}) + \beta_3 \text{Teacher\_Education}_i + \beta_4 \text{Parent\_Education}_i \\ &+ \beta_5 \text{Locale} + \beta_6 I(\text{School}_i = \text{Private}) \times \text{Time}_{ij} + \beta_7 \text{Teacher\_Education}_i \times \text{Time}_{ij} \\ &+ \beta_8 \text{Parent\_Education}_i \times \text{Time}_{ij} + \beta_9 \text{Locale}_i \times \text{Time}_{ij}. \end{aligned}$$

Given that a nontrivial amount of the data was missing, we performed multiple imputation with chained equations using the mice package in R [5]. The data were imputed using the predictions from linear, logistic, proportional odds, and multinomial models for continuous, binary, ordinal, and categorical data respectively. One hundred imputed datasets were created and the GLS model was fit to each one simultaneously using cluster computing. This reduced the computational time from approximately 8 hours to 5 minutes. The estimates were then combined according to Rubin's multiple imputation rules [6]. Any model that did not converge was discarded from the analysis. A forest plot of the point estimates and 95% confidence intervals is shown:



Based on the forest plot above, we see a strong positive time trend in the cohort's math scores, as expected. Children who have parents with a higher education perform significantly better in kindergarten, but have significantly less improvement in their scores over time. Schools that are located in rural areas, towns, or suburban areas all have significantly higher math scores at baseline compared to urban schools, but we see no significant differences in math score improvement over time.

Discussion

The results of this analysis have important policy implications. For example, future work is needed to uncover why public schools have more inequality in math scores, and what interventions can be used to reduce these education gaps. In our longitudinal modeling, we observed that having a highly-educated parent is significantly associated with higher scores in kindergarten. Schools and policymakers may attempt to identify the specific factors that are driving this association and provide all students with the same resources.

References

[1] K. Fiscella and H. Kitzman. Disparities in academic achievement and health: The intersection of child education and health policy. *Pediatrics (Evanston)*, 2009.

[2] National Center for Education Statistics. Early childhood longitudinal studies program. https://nces.ed.gov/ecls/.

[3] StatsDirect. Gini coefficient of inequality. https://www.statsdirect.com/help/nonparametric_methods/gini_coefficient.htm.

[4] M.S. Handcock and M. Morris. *Relative Distribution Methods in the Social Sciences.* Springer,New York, ISBN 0-387-98778-9, 1999.

[5] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

[6] R.J. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Wiley, 3rd edition, 2020.