Using Student Test Data to Predict Socio-Economic Status
By Emily Shepherd
DSCI 605

Project Rationale:
A common criticism of using standardized test data to evaluate both students and schools is that the test scores often reflect much more than content knowledge. Particularly, the overwhelming concern is that socioeconomic issues have a large impact on these scores. My goal with this project is to use test scores with other factors to try to predict a student's socio-economic status. In schools, we often use school lunch status to determine socio-economic level.

Part 1 Exploratory Data Analysis:
Initially, I created several visualizations to determine by cursory inspection whether a relationship between test scores and lunch status appear to have some association. The plots in fig. 1 and fig 2 explore the relationship between test scores. The color of each point corresponds to the whether the individual student receives free/reduced lunch.

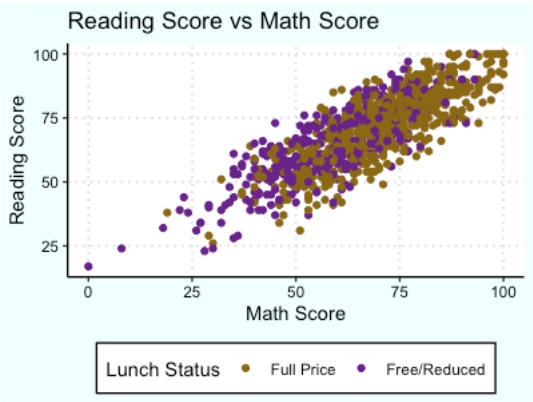


Fig. 1: Reading Scores vs Math Scores

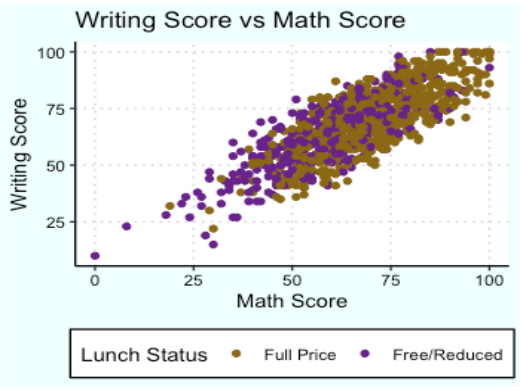


Fig. 2: Writing Scores vs Math Scores

From the visualizations in figures 1 and 2, it appears that there is a large percentage of high scores for students who do not qualify for free/reduced lunch. Likewise, students who scored lower on the three different tests were more likely to receive free/reduced lunch.

	Median Math Score	**Median Reading Score**	**Median Writing Score**
Fully Paid Lunch	69	72	72
Free/Reduced Lunch	60	65	64

Table 1: Median Scores Grouped by Lunch

Since the math scores serve as a strong explanatory variable for both writing and reading scores, when looking at different subgroups in relation to test scores, the preliminary analysis will focus solely on the math scores.

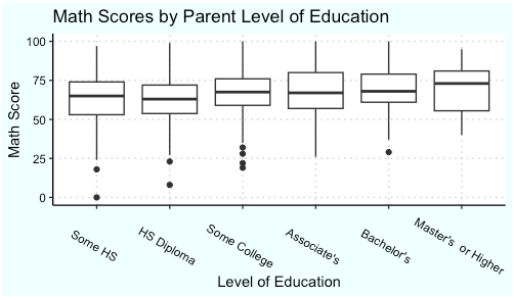
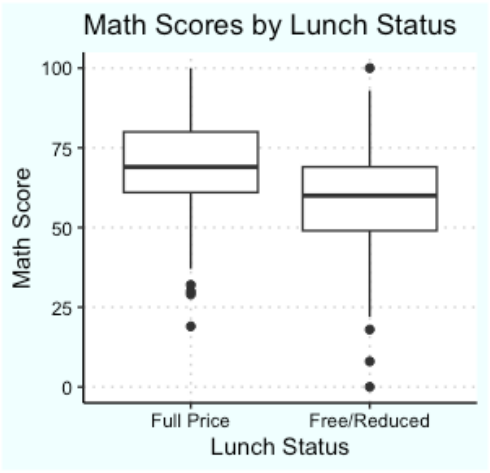


Fig. 3: Distribution of Math Scores by Parent Education



Fig, 4: Distribution of Math Scores by Lunch Status

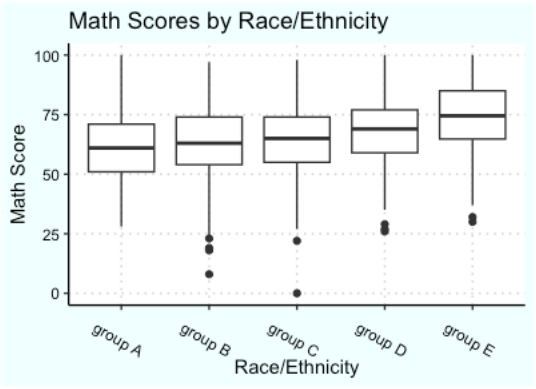


Fig. 5: Distribution of Math Scores by Race/Ethnicity

Figure 3 shows the median is about the same for all groups. However, the spread is much smaller for students whose parents hold college degrees, especially with bachelor's degrees and master's degrees. From the facet grid in Figure 6, it shows there are far fewer students who have attained bachelor's or master's degrees. This could account for the smaller spreads.

In figure 4, we can see that students who pay full price for lunch, have a higher median math score. Table 1 shows the median math score for students with full-priced lunch is 69; whereas, the median math score for students who qualify for free/reduced lunch is 60. Figure 5 shows the distribution of math scores grouped by race/ethnicity. The data does not specify which group corresponds to which racial classification. The median scores for groups D and E are slightly higher than the others. However, there does not seem be much variation among the five groups.

Fig. 6: Distribution of Math Scores faceted by Parent Education and Ethnicity/Race

Part 2 Creating a Logistic Model:

The variables math score, reading score, and writing score were used to create the model. Parental education level and race/ethnicity were not used as predictive variables, because there are real world associations between these variables and socioeconomic status. Also, from performing the initial exploratory analysis, it appeared there were not huge differences in the distributions based on race/ethnicity and parental level of education. Before creating the model, I split the data randomly, so that 70% of the data would be used to train the model, and the remaining 30% would be used to test the model. The summary of the model can be seen in Table 1.

Coefficients	Estimate	Std Error	z-Value	P-value
math.score	-0.07995	0.0111	-7.231	4.81e-13
reading.score	0.0680	0.0213	3.195	0.0014
writing.score	-0.0399	0.0195	-2.043	0.0410

Table 2: Summary of the Logistic Model Predicting Socioeconomic Status from Math, Reading, and Writing Scores

Part 3 Analysis of Model:
Using the test data to predict the lunch status resulted in the confusion matrix shown in Table 3. For students who paid full price for lunch, the model predicted 174 correctly and 62 incorrectly. Of the students who received free/reduced lunch, the model only predicted 20 correctly, and misidentified 45.

	Predicted Full Price	**Predicted Free/Reduced**
Actual Full Price	174	62
Actual Free/Reduced	45	20

Table 3: Confusion Matrix for the Logistic Model Predicting Socioeconomic Status

Sensitivity	0.7373
Specificity	0.6923
Pos Pred Value	0.8969
Neg Pred Value	0.4206
Prevalence	0.7841
Detection Rate	0.5781
Detection Prevalence	0.6445
Balanced Accuracy	0.7148

Table 4: Summary of Predictions on the Testing Data

Overall, the model accurately predicted the lunch status for 72.8% of the students.

Fig. 7: Graph of the Logistic Regression Curve for the Model

From Figure 7, we can see the prediction curve. The points represent the actual values, 0 for Full Price and 1 for Free/Reduced, in our training set. Any prediction of 0.5 or higher, would be

mapped to 1, meaning Free/Reduced. From looking at the graph, a combined score for Math, Reading, and Writing of about 165 would be predicted to belong to a student who received free/reduced lunch.

Part 4 Conclusion:

In general, the model had some success predicting socioeconomic status as measured by qualification for free/reduced lunch. While this data represented scores and demographics from students at one school, the methods might be more accurate for statewide data and limited to a single grade level. For example, creating a model based on iLearn scores for a random sample of fourth grade students in Indiana might be more accurate in predicting which students qualify for free/reduced lunch. By being able to do this, it would establish that standardized tests are indicators of factors other than merely academic performance.