Discovering which socioeconomic variables have highest influence on average ACT Score

Jack Neton

DATA 5100

18 Oct 2025

**Introduction**

College admissions testing is primarily evaluated using two standardized tests:

* American College Test (ACT) test, or
* Scholastic Assessment Test (SAT).

Our study focuses on the ACT. The ACT has four multiple choice sections: English, Math, Reading, and Science. Each section is scored from 1-36; a student’s score is the average across all four sections.

College admissions testing has been a curious question from a socioeconomic standpoint. Do certain household or regional wealth factors affect accessibility to test prep, or a district’s overall ability to prepare their students?

This study aims to fill that gap and answer the question of which socioeconomic variable affects ACT score the most.

**The Data**

Two data sets were used in this study.

The [EdGap](https://github.com/brian-fischer/DATA-5100/blob/main/EdGap_data.xlsx) data set contained 7 columns of data. This set contained all the socioeconomic variables and the average ACT score.

The [Schools Information](https://www.dropbox.com/scl/fi/fkafjk8902sq8ptxh94r2/ccd_sch_029_1617_w_1a_11212017.csv?rlkey=gucrdz5f6e38bezz2y3yalxbw&e=1&dl=0) data set contained 65 columns of data. This set contained information of the schools, mainly geographical location and school type.

One of the most critical variables was the School ID variable, found in both sets, as it was used to join the two data sets together.

**Analysis Methods**

Simple analysis of the EdGap data were completed first to make sure that the data were suitable to answer the question. This was done with a pair plot (with and without best fit lines) to see correlation between variables. The pair plots signified a relationship between the socioeconomic variables and the ACT score, meaning I could proceed.

I then selected relevant subsets of the school information data set, reducing the size from 65 columns to 7. This made it easier to complete a left join on School ID of the two data sets. A left join was used to retain all the relevant socioeconomic variables in the EdGap data set.

A quick geographical plot showed the limits of the study after the join; several states had been lost due to the join.

After cleaning NA values and removing rows with no ACT score, we used an imputer to replace missing values. Once the number of missing values was zero, we could proceed with proper analysis.

I then defined predictor variables of median income, college education of parents, married parents percentage, unemployment rate, percentage on school lunch, state, and charter.

The following methods were done individually on median income, college educated parents, and percentage of school lunch:

* A scatter plot of the regression line
* An OLS regression
* R-Squared, Root Mean Squared Error, and Mean Absolute Error
* A residual scatter plot
* A quadratic model scatter plot
* A fit test of the quadratic model
* Anova test to compare regression models

After individual analysis of the variables at hand, a quick OLS regression was done for state and school type.

Next, multiple regression was carried out on the predictor variables, except state and charter. This yielded the need for a reduced model, and multiple regression analysis was conducted for unemployment rate, percent of college educated parents, and percent on school lunch.

I then scaled the model above to compare the accuracy of the models.

**Results**

School lunch had the highest weight as a predictor of average ACT score. Consistently, school lunch was more statistically significant in all the tests completed; it had highest R-squared and lowest p-values.

Interestingly, state and school type did not have statistically significant weight on the average ACT Score.

After scaling, I saw that my original model sufficed for determining statistical significance of the variables in the study. Scaling helped confirm the magnitude that school lunch has on average ACT score compared to the other predictor variables.

**Conclusion**

Surprisingly, the percentage of students on school lunch was the highest predictor of average ACT score. I walked in to this study assuming it would be State or median income. It was interesting to find that, on a granular level with something like school lunch, there’s more than meets the eye for predicting average ACT score.