DATA5100 – Education Project

Hannah Belville

Socioeconomic Determinants of School Performance

Abstract

The ACT (American College Testing) exam is a standardized test commonly used in the United States for college admissions and assessment of academic readiness. The score reflects student performance across four subject areas—English, mathematics, reading, and science—and serves as a benchmark for both colleges and policymakers when evaluating school performance. This study investigates whether school performance, measured by the average ACT composite score, can be predicted from socioeconomic factors. Using a data science workflow, several publicly available datasets from EdGap.org, the National Center for Education Statistics (NCES), and U.S. Census EDGE Income-to-Poverty Ratio (IPR) estimates were combined and analyzed. After data cleaning, imputation, and exploratory analysis, a multiple linear regression model was fitted using predictors such as median income, parental education level, family structure (married vs. divorced parents), unemployment, and free/reduced lunch rates. Results show that higher community income and adult education levels are associated with higher ACT performance, while a higher proportion of students receiving free or reduced-price lunches significantly predicts lower average scores. These findings highlight the strong influence of socioeconomic conditions on standardized testing outcomes and underscore the role of broader social inequality in shaping educational achievement.

1. Introduction

The central question is:  
Can socioeconomic indicators such as income, education level, and family structure meaningfully explain variations in high school ACT performance across the United States?

Socioeconomic factors play a central role in understanding variation in standardized test outcomes. Income, parental education, employment status, and family structure can affect access to educational resources, extracurricular opportunities, and test preparation. Communities with higher socioeconomic status often benefit from stronger school funding, greater parental involvement, and more stable learning environments—all of which contribute to higher test performance.

2. Theoretical Background

Educational performance has long been understood as a function not only of individual ability and school quality, but also of broader socioeconomic conditions. These factors influence academic achievement through multiple pathways, ranging from access to advanced coursework and test preparation to the stability of home environments and expectations for college-level education.

Socioeconomic status is typically understood as a combination of economic, educational, and social resources. In this analysis, socioeconomic status is captured using several indicators:

* Household income: Reflects economic capital and determines access to educational materials, private tutoring, and outside learning opportunities.
* Parental education level: Captures human capital, as parental education can influence a child’s academic support at home.
* Family structure: Serves as a proxy for social stability and the availability of support systems affecting student engagement and performance.
* Unemployment rate: Provides context for community-level economic distress, which can influence both school funding and student stress levels.
* Free/reduced lunch rate: Measures economic disadvantage at the student level and provides more granular detail than income alone.
* Income-to-Poverty Ratio (IPR): Standardizes community income relative to federal poverty thresholds, contextualizing local wealth disparities.

By integrating datasets from NCES, EdGap.org, and EDGE IPR data, this study uses a quantitative modeling approach to evaluate how these indicators jointly predict ACT performance. The analysis builds on theoretical models of educational inequality, particularly the idea that academic outcomes reflect structural socioeconomic conditions rather than only school-level differences.

3. Data and Methods

Datasets were merged using the NCESSCH school ID, a unique identifier assigned by NCES. Data cleaning steps included type standardization, high school filtering, invalid value replacement, duplicate removal, and Iterative Imputer–based missing value imputation. Outliers were detected using the interquartile range (IQR) method, and continuous variables were scaled for consistency.

Exploratory analysis revealed strong relationships between income, education, and ACT performance. Schools with higher community education levels and income had higher ACT averages. Correlation analysis, as shown in Figure 1 below, found ACT score positively correlated with median income (*r* ≈ 0.45) and parents with a college degree (*r* ≈ 0.44), and negatively correlated with free/reduced lunch (*r* ≈ –0.77).

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1: Correlation Matrix of Predictor Variables

4. Computational Results

The regression model explained approximately 72% of the variance in ACT performance (R² ≈ 0.72).

While correlation coefficients (*r*) were used for exploratory pairwise relationships, predictor importance in the regression model is determined by standardized coefficients (β). These coefficients, shown in Figure 2 below, reflect each variable’s effect on ACT scores after controlling for all other predictors.

A graph of a graph of a red line

AI-generated content may be incorrect.

Figure 2: Numerical Predictors vs Average ACT Scores

Among standardized predictors:

* IPR level (β ≈ 29.40) and median household income (β ≈ 4224.13) had the strongest positive associations with ACT scores.
* Free/reduced lunch rate had the largest negative coefficient (β ≈ -0.07), consistent with its role as a proxy for economic disadvantage.
* Family structure, adult education level, and unemployment rate had smaller but meaningful effects.

The Income-to-Poverty Ratio (IPR) added explanatory power by contextualizing income relative to local poverty thresholds and helped reduce ambiguity among economic predictors. Residual analysis confirmed that model assumptions were reasonably met, and high-leverage points (from very small or affluent schools) had limited influence on estimates.

5. Discussion and Conclusion

The results confirm a robust link between socioeconomic advantage and academic performance. Schools in communities with higher median income and adult education levels consistently report higher ACT scores, while higher rates of free/reduced lunch are associated with lower scores. These findings align with research showing that standardized test performance reflects both instruction quality and the social and economic context of students.

Limitations:

* The datasets used different geographic and organizational units. ACT performance is reported at the school level, whereas many socioeconomic variables come from census tract or district-level estimates. This mismatch can introduce ecological inference errors, as tract-level averages may not fully represent each school’s student population.
* Census-derived indicators such as IPR are subject to sampling error, particularly in small or rural areas.
* Differences in data collection years introduce temporal uncertainty, and imputation assumptions may affect variable relationships.

Despite these limitations, the model explains over two-thirds of ACT variance, highlighting the strong influence of socioeconomic conditions. Addressing test score disparities require interventions beyond the school, including broader social and economic policies to reduce inequality.

6. References

Belville, H. (2025). *Education Analysis Notebook (Education.ipynb).* GitHub repository: <https://github.com/belvilleh/education>

EdGap.org. (2023). *Education and Socioeconomic Data by School District.* Retrieved from https://edgap.org

National Center for Education Statistics (NCES). (2023). *Common Core of Data (CCD) and ACT Score Reports.* U.S. Department of Education. Retrieved from https://nces.ed.gov/ccd/

U.S. Census Bureau. (2023). *Education Demographic and Geographic Estimates (EDGE) Program – Income to Poverty Ratio (IPR) Data.* Retrieved from https://www.census.gov/programs-surveys/edge.html