**School Performance and Socioeconomic Predictors**

## **Abstract**

**Purpose**: This project looks at whether we can predict school performance (measured by average ACT scores) using socioeconomic factors.

I followed a standard data science workflow throughout: loading data → cleaning → merging datasets → creating new features → exploring the data → building models → checking diagnostics → exporting results.

**Data sources**: EdGap dataset and CCD school dataset (provided in class).

For this project, I analyzed how socioeconomic factors affect school-level academic performance across the United States. I worked with data from EdGap and the Common Core of Data (CCD) datasets to look at the relationship between socioeconomic status (measured by the percentage of students on free or reduced-price lunch) and average ACT scores. I used multiple regression with robust standard errors and ran several diagnostic checks. My results showed a pretty strong positive relationship between higher SES and higher ACT scores, and this relationship held up even after I controlled for state-level differences and removed some outlier schools.

## **Introduction**

Educational achievement in the U.S. often reflects socioeconomic disparities - it's a well-known issue. Schools in wealthier areas typically have better resources, lower student-teacher ratios, and higher test scores. For this study, I wanted to investigate whether socioeconomic status can actually predict school-level performance on the ACT exam.

I combined two datasets:

* **EdGap data**: includes socioeconomic indicators like unemployment rate, adult education levels, and median household income
* **Common Core of Data (CCD)**: provides school-level information and characteristics

By merging these datasets using the school ID (NCESSCH), I aimed to figure out how strongly SES relates to academic outcomes. I also made sure to run proper diagnostics and sensitivity checks to make sure my results were solid.

## **Results - Regression Summary**

* **Model**: I used OLS regression to predict avg\_act from ses\_index (where higher values = wealthier communities). I used robust (HC1) standard errors to account for heteroskedasticity.
* **Key finding**: The ses\_index coefficient came out to **8.2958** (robust SE = **0.086**, p < 0.001).
* **Intercept**: **23.6786**
* **Model fit**: Adjusted R² = **0.605**, which means the model explains about 60.5% of the variation in average ACT scores across schools in my sample.

All of these estimates use heteroskedasticity-robust standard errors (HC1).

## **Theoretical Background**

Socioeconomic status has been linked to academic achievement for a long time - mainly through things like access to resources, parental education levels, and community support. Research from the National Center for Education Statistics (NCES) and other educational studies consistently shows that students from higher-income families do better on standardized tests like the ACT and SAT.

This pattern usually comes from disparities in school funding, availability of advanced classes, and access to test prep resources. For this study, I measured SES using variables like median household income, unemployment rate, and the percentage of students qualifying for free or reduced lunch. This approach fits within established research showing that higher SES communities tend to provide better environments for academic success.

## **Methodology**

I followed a reproducible workflow throughout this project: data loading → cleaning → merging → feature engineering → exploratory analysis → modeling → diagnostics.

### **Datasets**

I used two main datasets:

* **EdGap**: county-level socioeconomic indicators (unemployment rate, adult education levels, median household income)
* **Common Core of Data (CCD)**: school-level characteristics and ACT averages

### **Data Processing**

I merged the datasets using NCESSCH (the school ID code). For data cleaning, I handled missing values, converted data types where needed, and standardized column names to make everything consistent.

I created a new variable called ses\_index by taking the negative of the percentage of students eligible for free/reduced-price lunch. This way, higher values indicate more affluent schools, which makes the interpretation easier.

### **Statistical Analysis**

I used Ordinary Least Squares (OLS) regression to estimate:

avg\_act = β₀ + β₁(ses\_index) + ε

I applied robust (HC1) standard errors to correct for heteroskedasticity. For diagnostics, I looked at Q-Q plots, residual plots, and Cook's D to detect influential observations.

For sensitivity checks, I added state fixed effects and re-ran the model after removing influential outliers to make sure my results were stable.

## **Computational Results**

My baseline OLS model gave me these results:

* **Coefficient for ses\_index**: 8.2958 (robust SE = 0.086, p < 0.001)
* **Intercept**: 23.6786
* **Adjusted R²**: 0.605

This means that a one-unit increase in the SES index corresponds to about an 8.3-point increase in average ACT scores, holding everything else constant.

The diagnostics showed mild heteroskedasticity but no major violations of the linear model assumptions. When I added state fixed effects, the SES coefficient stayed pretty much the same (β₁ ≈ 8.30). After I removed outliers using Cook's D, I got similar results (β₁ ≈ 8.58).

These findings suggest that the relationship between SES and school performance is pretty robust and holds up across different model specifications.

## 

## **Discussion**

My analysis supports the idea that socioeconomic conditions significantly predict school-level ACT performance. Schools in more affluent areas tend to achieve higher test scores, which reflects broader advantages like better resources, teacher retention, and family support networks.

The relatively high adjusted R² (about 0.605) suggests that SES explains a large chunk of the variation in school performance - which is honestly a reminder of how deeply educational inequality is tied to economic inequality.

I also had access to the Hospital General Information and 2020 Census datasets, which provided some useful context. Regions with better healthcare access and stronger demographic indicators of stability often overlap with higher SES areas, which reinforces the educational disparities I found. While I didn't merge these datasets into my regression model, they help show how multifaceted school performance outcomes really are.

## **Conclusions and Limitations**

### **Conclusions**

Overall, my results show that socioeconomic disparities are a strong and measurable predictor of educational performance in U.S. schools.

Socioeconomic status (measured by percent free/reduced lunch) is strongly associated with school-level ACT performance in this dataset. The relationship is both statistically significant and practically meaningful - even modest improvements in SES correspond to real gains in average ACT scores.

The model explains a large share of the variation in ACT scores (adj. R² ≈ 0.605), suggesting that SES and related factors are important predictors at the school level.

### **Limitations**

**Observational design**: This analysis is correlational, so I can't claim any causal effects without stronger methods like panel data, instrumental variables, or quasi-experiments.

**Omitted variables**: There are probably other factors affecting the SES-ACT relationship that I didn't include - things like school resources, teacher quality, student selection processes, and other community factors.

**Residual distribution**: My diagnostic tests (Omnibus, Jarque-Bera) showed non-normal residuals with heavy tails. I used robust standard errors (HC1) to deal with heteroskedasticity, but the non-normality means I should be cautious about predictions at the extremes.

**Measurement issues**: My SES proxy (pct\_free\_lunch) isn't a perfect measure of socioeconomic context. Any measurement error would affect the accuracy of my coefficient estimates.

## **References**

* U.S. Department of Education, Common Core of Data (CCD), 2017.
* EdGap Data (provided via DATA 5100 course repository).
* Hospital General Information Dataset, U.S. Department of Health & Human Services.
* 2020 U.S. Census Disclosure Avoidance System Dataset.