Education Project Report

Alex Song
DATA 5100

Abstract

This project analyzes relationships between high-school ACT averages and a set of school-level socioeconomic and funding variables. Using data from the EdGap and School Information datasets provided in class and the 2017 Public Elementary–Secondary Education Finance dataset from census.gov, the study explores how these factors relate to educational performance. After cleaning, merging, and imputing missing numeric values with an *Iterative Imputer*, regression models were estimated to assess explanatory power. Socioeconomic indicators—median household income, adult education, unemployment, and free/reduced-lunch percentages—accounted for most of the variation in ACT scores ($R^2 \approx 0.63$). Funding variables such as per-pupil expenditures and instructor salaries showed statistically significant but far weaker relationships ($R^2 \approx 0.004–0.034$) and were highly collinear ($r \approx 0.89$). Overall, socioeconomic context appears to be the most reliable predictor of average ACT performance.

1. Introduction

Educational performance, often measured by standardized assessments such as ACT scores, is influenced by both school resources and community conditions. This analysis asks:

Are differences in high-school ACT scores better explained by socioeconomic characteristics or by school funding levels?

To answer this question, the study merges data on school-level ACT outcomes, socioeconomic indicators, and district-level financial statistics. The goal is to identify which predictors offer the strongest and most consistent explanation of average ACT results. The following sections describe the theoretical framework, data preparation, analytical methods, key findings, and conclusions.

2. Theoretical Background

Linear regression provides a straightforward way to estimate how an outcome changes with predictors while controlling for other variables. These estimates measure **association**, not **causation**. When predictors are strongly correlated—such as per-pupil expenditure and instructor salary—it becomes difficult to isolate their individual effects.

Socioeconomic factors such as local income, parental education, and unemployment shape students' preparation for standardized testing. Financial variables also vary across districts, yet they are often intertwined with community wealth. Importantly, the socioeconomic variables are available at the **school** level, while the financial measures are aggregated at the **district** level. This mismatch can introduce bias: district averages may obscure variations among individual schools. Therefore, interpretations of combined analyses should be made cautiously.

---

3. Methodology

3.1 Data Preparation

Three datasets were merged so that each observation represented one high school with its ACT average, socioeconomic attributes, and available district-level funding information:

1. **EdGap** – School outcomes and socioeconomic indicators
2. **School Information** – Metadata on school classification
3. **2017 Education Finance (Census.gov)** – District funding data

Invalid ACT values were removed, and only schools labeled "High" were retained. Column names were standardized for clarity (e.g., *PPCSTOT* → *PerPupil_Total_Expenditure*). Because the finance data were reported at the district level, some mismatches occurred when merging with school-level data.

Missing data were minimal (< 1 %) for socioeconomic indicators but higher (~ 5–6 %) for funding measures. Numeric gaps were filled using an **Iterative Imputer**, preserving sample size and relationships among variables.

3.2 Exploratory Analysis

Summary statistics, correlations, and scatterplots (linear + quadratic fits) were used to explore distributions. Funding indicators—*PerPupil_Total_Expenditure* and *PerInstructor_Salary*—were highly correlated (r ≈ 0.89), indicating overlapping information.

3.3 Modeling Strategy

Simple linear regressions tested each predictor's relationship with ACT scores. Median income was modeled with both linear and squared terms to assess curvature. Multiple-regression models combined socioeconomic predictors to evaluate collective explanatory power.

Model performance was evaluated using **R^2**, **mean absolute error (MAE)**, and **root mean squared error (RMSE)**. Diagnostic checks confirmed no severe violations of linear-model assumptions.

4. Computational Results

After cleaning, the dataset included approximately 7,200 high schools (original ≈ 7,900).

- **Socioeconomic Models:** Median household income alone explained ≈ 21 % (R^2 = 0.211) of variation in ACT scores. Adding a squared term marginally improved fit, suggesting a slightly curved, leveling relationship at higher incomes.
- **Funding-Only Models:** Per-pupil spending explained < 1 % of variation (R^2 = 0.004); instructor salary explained ≈ 3 % (R^2 = 0.034).
- **Combined Socioeconomic Models:** Including multiple socioeconomic variables increased explanatory power to ≈ 63 % (R^2 = 0.628).
- **Combined (Socioeconomic + Funding):** Adding funding variables produced minimal additional improvement.

Table 1 – Model Performance Summary

Model Type	**Predictors**	**R^2**	**MAE**	**RMSE**
Funding only	Expenditure, Salary	0.03	2.4	3.1
Socioeconomic only	Income, Education, Unemployment, Lunch Rate	0.63	1.8	2.2
Combined	All variables	0.64	1.7	2.1

5. Discussion

The analysis demonstrates that a few well-chosen socioeconomic indicators—household income, adult education, and student economic need—account for most of the variation in

ACT performance across schools. Once these variables are included, the explanatory value of funding measures becomes minimal.

This result aligns with previous educational research, including large-scale studies such as the **Coleman Report**, which found that community socioeconomic status strongly influences student achievement. The high correlation between funding and income ($r \approx 0.89$) suggests that wealthier communities naturally allocate more resources, blurring distinctions between “money effects” and “context effects.”

Possible alternative explanations include unobserved factors such as teacher quality, curriculum rigor, and parental involvement. Furthermore, since funding data are aggregated at the district level, intra-district disparities are not captured, limiting causal interpretation.

---

6. Conclusion

Socioeconomic context is the most consistent and powerful predictor of high-school ACT scores in these merged data. Financial indicators—per-pupil expenditure and instructor salary—show some association with scores but are highly collinear and closely tied to community wealth. Consequently, once socioeconomic differences are accounted for, funding variables add little explanatory power.

These findings imply that **addressing socioeconomic disparities** may yield greater educational improvement than isolated increases in school funding. Future research could incorporate student-level or longitudinal data to test causal mechanisms and capture within-district variation.