Education Project: Socioeconomic Factors Affecting School Performance

Abstract

This study examines how socioeconomic and school level factors predict ACT scores in U.S. high schools using regression analysis on merged national datasets. Among all predictors tested, the percentage of students receiving free or reduced lunch was the strongest indicator, explaining over 60% of variation in ACT performance. These results highlight how school level economic disadvantage remains a key driver of educational inequality.

Introduction

Educational outcomes in the United States continue to reflect deep socioeconomic inequalities. Standardized tests such as the ACT provide a measurable indicator of academic performance, yet scores vary widely across schools due to structural and economic disparities. Understanding these patterns is essential for identifying the factors that most strongly influence educational opportunity.

Prior research shows that family income, employment stability, and school funding strongly influence educational outcomes by shaping access to academic resources and preparation opportunities. Examining how these conditions vary across schools provides insight into the connection between socioeconomic context and student performance.

The data used in this study combine information from three sources: the EdGap dataset, which includes average ACT/SAT scores and socioeconomic variables such as median income and unemployment rate; and two datasets from the U.S. Department of Education's Common Core of Data (CCD), which provide school-level details such as teacher counts, school type, and free or reduced lunch participation. These datasets were cleaned, merged using a common school identifier, and prepared for analysis.

This study aims to determine which socioeconomic or school-level variable best predicts average ACT performance among U.S. high schools. By quantifying these relationships through regression analysis, the study contributes to understanding how educational inequality manifests within the American school system and where interventions may be most effective.

Theoretical background

Theoretical frameworks in education research suggest that access to financial, social, and academic resources determines the quality of learning environments and student outcomes. Students from lower-income households often face limited access to educational materials, advanced coursework, and extracurricular learning opportunities, which collectively affect standardized test performance.

At the school level, socioeconomic disadvantage can be reflected in indicators such as unemployment rate, median household income, and the percentage of students eligible for free or

reduced lunch. These variables capture different dimensions of economic stability, community context, and family resources. Similarly, institutional characteristics such as the number of teachers may represent structural capacity but not necessarily instructional quality. Understanding how these variables interact with student performance provides a data-driven way to examine the mechanisms of educational inequality.

This study applies linear regression analysis to test the strength and direction of relationships between each predictor and average ACT scores. Regression modeling is appropriate for this context because it quantifies how much variation in academic outcomes can be explained by differences in school-level socioeconomic factors, allowing for direct comparison among predictors.

Methodology

This study uses a merged dataset combining information from the EdGap dataset, which includes ACT/SAT scores and socioeconomic variables for U.S. high schools, with two Common Core of Data (CCD) datasets containing school-level characteristics such as location, charter status, teacher count, and percentage of students receiving free or reduced lunch. The EdGap data include variables such as median household income, unemployment rate, and percentage of students from married families, capturing socioeconomic conditions at the school and district levels.

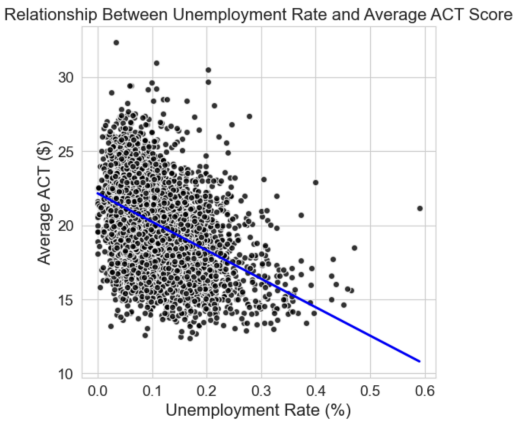
Data preparation involved removing unnecessary columns and retaining only variables relevant to the analysis, including school ID, unemployment rate, median income, percent of married families, percent of students receiving free or reduced lunch, charter status, number of teachers, school year, and average ACT score. The datasets were merged using a left join on the school identifier to ensure all records from the EdGap dataset were preserved. Missing values were handled using an Iterative Imputer from scikit-learn, which estimates missing entries based on correlations among other features.

Following data cleaning and merging, a series of simple linear regression models were fitted using the ols() function from the statsmodels library in Python. Each predictor variable (unemployment rate, free/reduced lunch percentage, and number of teachers) was modeled individually against average ACT score to isolate its effect. Model performance was evaluated using the coefficient of determination (R^2), p-values for statistical significance, and error metrics such as mean squared error (MSE) and mean absolute error (MAE). Residual plots and scatterplots were used to assess linearity and model fit.

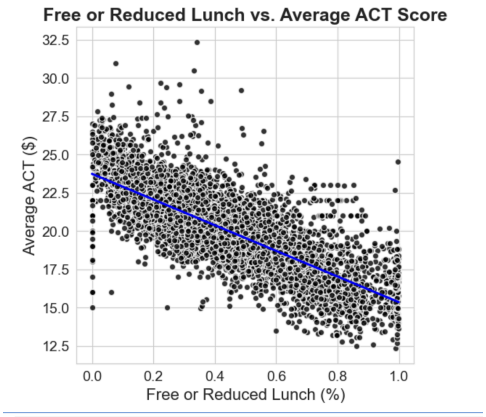
Computational Results

The first model examined the relationship between the unemployment rate and average ACT scores. The regression plot revealed a clear negative trend—schools located in areas with

higher unemployment rates tended to report lower average ACT scores. The fitted linear model confirmed this relationship, with a statistically significant negative coefficient ($\beta = -19.21$, $p < 0.001$). The coefficient of determination ($R^2 = 0.188$) indicates that unemployment rate explains about 19% of the variance in ACT performance, suggesting a moderate association. While meaningful, the wide scatter of data points around the regression line indicates that additional factors likely influence ACT outcomes beyond the local unemployment rates.



The second model examined the relationship between the percentage of students eligible for free or reduced lunch and average ACT scores. The regression results revealed a strong and statistically significant negative association ($\beta = -0.09$, $p < 0.001$), indicating that schools with higher proportions of students qualifying for free or reduced lunch tended to have lower average ACT scores. The model's R^2 value of 0.614 shows that this single variable explains over 60% of the variation in ACT scores, far more than the unemployment rate model. This finding suggests that the free/reduced lunch variable is a powerful predictor of academic performance, as it directly reflects school-level economic disadvantage.



OLS Regression Results

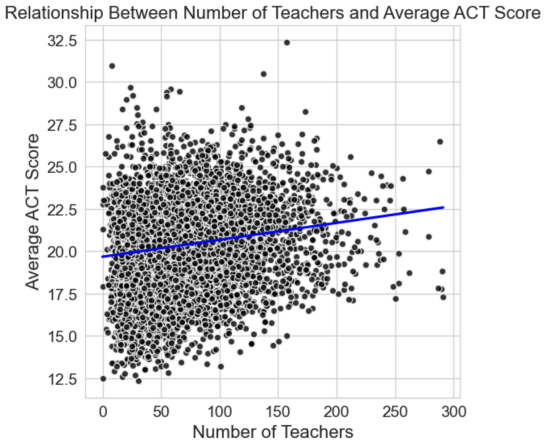
Dep. Variable:	average_act	R-squared:	0.614
Model:	OLS	Adj. R-squared:	0.614
Method:	Least Squares	F-statistic:	1.149e+04
Date:	Sun, 19 Oct 2025	Prob (F-statistic):	0.00
Time:	16:22:53	Log-Likelihood:	-13461.
No. Observations:	7227	AIC:	2.693e+04
Df Residuals:	7225	BIC:	2.694e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.7429	0.037	641.759	0.000	23.670	23.815
percent_lunch	-8.3902	0.078	-107.187	0.000	-8.544	-8.237

Omnibus:	842.255	Durbin-Watson:	1.472
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2849.644
Skew:	0.582	Prob(JB):	0.00
Kurtosis:	5.848	Cond. No.	5.02

The third model explored the relationship between the number of teachers at a school and average ACT scores. The regression results showed a statistically significant but weak positive

relationship ($\beta = 0.01$, $p < 0.001$), suggesting that schools with more teachers tend to have slightly higher ACT scores. However, the model's R^2 value of 0.03 indicates that teacher count explains only about 3% of the variation in ACT performance. This limited explanatory suggests that while larger schools may employ more teachers, teacher count alone does not meaningfully predict student achievement.



OLS Regression Results

Dep. Variable:	average_act	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	221.3
Date:	Sun, 19 Oct 2025	Prob (F-statistic):	2.56e-49
Time:	16:23:06	Log-Likelihood:	-16791.
No. Observations:	7227	AIC:	3.359e+04
Df Residuals:	7225	BIC:	3.360e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	19.6757	0.051	385.992	0.000	19.576	19.776
teachers	0.0100	0.001	14.875	0.000	0.009	0.011

Omnibus:	50.453	Durbin-Watson:	1.206
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54.194
Skew:	-0.173	Prob(JB):	1.71e-12
Kurtosis:	3.245	Cond. No.	133.

Discussion

The analysis identified the percentage of students eligible for free or reduced lunch as the strongest predictor of average ACT scores, explaining over 60% of the variation in performance. Schools with higher proportions of students qualifying for free or reduced lunch often serve communities with fewer economic and educational opportunities, which can limit access to college preparation materials, tutoring, and other academic supports.

In contrast, unemployment rate and teacher count were statistically significant but explained far less variation in ACT outcomes. The unemployment rate reflects broader community-level conditions rather than the direct economic realities of students attending a specific school, which likely weakens its predictive power. Similarly, the number of teachers is more reflective of school size than of classroom quality or instructional support, explaining why it does not meaningfully predict performance.

While these findings are consistent with socioeconomic theory, they should be interpreted within the limits of the available data. After cleaning and merging, the dataset included schools from only about 20 U.S. states, which restricts how broadly the results can be generalized. Additionally, key school specific variables such as tutoring availability, school funding, and academic support services were not included, and data were limited to the 2016–2017 school year. Expanding this analysis to include more states, additional school-level indicators, and

multiple years of data could provide a more comprehensive picture of how socioeconomic conditions shape educational outcomes.

Overall, these results highlight that school-level measures of economic disadvantage, such as free and reduced lunch participation, provide a more accurate picture of student performance than regional or structural indicators. The findings reinforce the importance of addressing material inequality within schools to create more equitable educational opportunities.

Conclusion

This study set out to determine which socioeconomic and school-level factors best predict average ACT scores among U.S. high schools. The analysis found that the percentage of students eligible for free or reduced lunch is the most powerful predictor of academic performance, reflecting how economic inequality translates directly into measurable differences in achievement.

These findings highlight that challenges in academic performance are rooted not only in individual effort but in broader structural disparities. School-level measures of economic disadvantage capture the lived realities of students more effectively than broader regional indicators such as unemployment rate or school-level attributes like teacher count.

The results suggest that educational interventions should focus on addressing economic barriers within schools. Policies that expand access to free tutoring, test preparation, and classroom resources, especially for low-income students. This could help narrow performance gaps and ensure that standardized assessments reflect ability rather than financial background. Strengthening school funding and resource distribution would be a meaningful step toward promoting equity and supporting student success nationwide.

Education Project: Socioeconomic Factors Affecting School Performance

Sources

U.S. Department of Education, National Center for Education Statistics. (2017). Common Core of Data (CCD): Public Elementary/Secondary School Universe Survey, 2016–2017. U.S. Department of Education. https://nces.ed.gov/ccd

EdGap Project. (2017). EdGap: Educational Opportunity Gap Dataset (2016–2017) [Data set]. Retrieved from https://github.com/brian-fischer/DATA-5100/blob/main/EdGap_data.xlsx