Education Project: Socioeconomic Factors Affecting School Performance

Abstract

In this project, I used economic indicators and school characteristics to evaluate which variables best predict average ACT scores. The analysis draws on three datasets: two containing school-level information such as teacher counts, location, and charter status, and one ("EdGap") providing socioeconomic data including median income, family structure, and unemployment rate. These datasets were merged into a single file to explore how both economic and structural factors relate to academic performance. The goal of this analysis is to identify which variable serves as the strongest predictor of average ACT scores and to better understand how socioeconomic inequality is reflected in educational outcomes.

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, but scores often vary widely across schools due to underlying structural and economic differences. Understanding these patterns can help identify which factors most strongly influence educational opportunity.

This study uses a merged dataset combining information from the EdGap dataset, which contains 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 free/reduced lunch participation. The EdGap dataset includes variables such as median household income, unemployment rate, and the percentage of students from married families. Together, these data sources capture both socioeconomic and institutional dimensions of educational environments.

To prepare the data, I removed unnecessary columns and retained only variables relevant to the analysis (School id, rate_unemployment, percent_married, median_income, percent_lunch, charter, teachers, state, school year, school level, and average_act). The datasets were merged using a left join on the school identifier (id) so that all schools from the EdGap dataset, the main source containing ACT data, were preserved. Missing values were addressed using an "Iterative Imputer", which predicts missing entries based on relationships among other features, allowing for more accurate estimates compared to simple mean or median imputation.

After cleaning and merging, I used simple linear regression models to test the relationship between each predictor and average ACT performance. The aim of this study is to determine which variable is the strongest predictor of the average ACT score and contributes to a deeper understanding of how economic disadvantage and school-level structure influence academic achievement.

Theoretical background

Education Project: Socioeconomic Factors Affecting School Performance

Socioeconomic theory suggests that differences in income, employment, and access to resources shape the quality of education that students receive. These differences often appear in measurable school-level indicators such as unemployment rate, percentage of students receiving free or reduced lunch, and number of teachers. I chose these variables because they reflect different dimensions of opportunity: economic stability, family resources, and school staffing, all of which can influence academic performance.

To examine how these factors relate to ACT outcomes, I used "linear regression," a method well suited for testing relationships between continuous variables. Regression modeling helps identify not just the direction of a relationship (positive or negative) but also how strong it is. It produces key metrics such as R-squared, which shows how much of the variation in ACT scores is explained by each variable, and **e**rror values like the mean squared error (MSE) and mean absolute error (MAE), which measure how closely the model's predictions match actual results. These metrics help evaluate both the strength and accuracy of each model, allowing for meaningful comparisons across predictors and a clearer understanding of which factors matter most for explaining educational outcomes.

Methodology

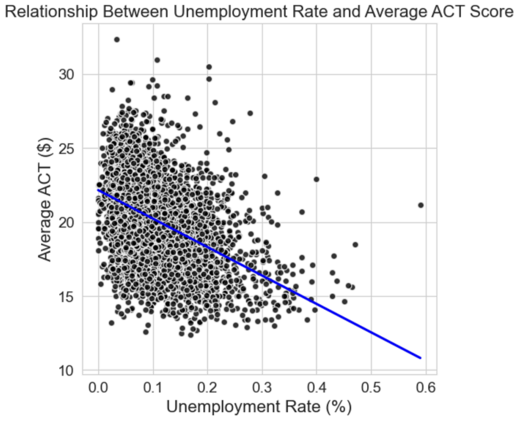
I conducted a series of simple linear regression analyses to evaluate which variables most strongly predict average ACT scores. Each predictor (unemployment rate, percentage of students receiving free or reduced lunch, and number of teachers) was modeled against the outcome variable Average ACT. This approach allowed me to isolate the effect of each factor and compare their predictive strength directly.

The models were implemented using the statsmodels library in Python, specifically the ols() function from statsmodels.formula.api. Each regression model generated a summary output containing key statistics, including the coefficient (direction and strength of association), p-value (statistical significance), and R-squared (proportion of variance in ACT scores explained by the predictor).

To assess model performance, I also calculated mean squared error (MSE) and mean absolute error (MAE) as measures of prediction accuracy. Additionally, residual plots were used to visually check for linearity and model fit. These steps ensured that the results were interpretable. I also created visualizations with the seaborn and matplotlib libraries to illustrate the relationships between variables and validate model assumptions.

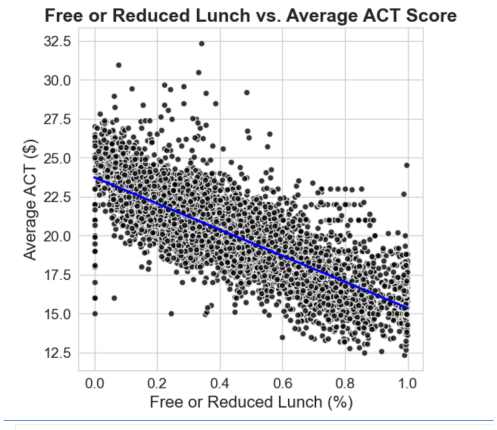
Computational Results

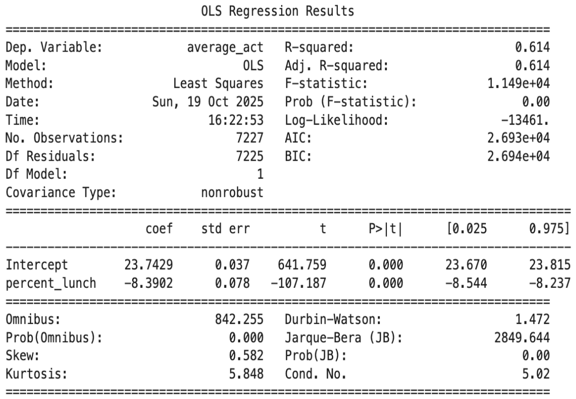
The first single-input model examined the relationship between unemployment rate and average ACT scores. A regression line was plotted to visualize the association. The scatterplot revealed a negative linear trend, indicating that higher unemployment rates were generally associated with lower ACT scores. However, the data points were widely dispersed around the regression line, suggesting only a moderate model fit.



To quantify this relationship, a simple linear regression model was fitted using unemployment rate as the predictor variable and average ACT score as the outcome. The model produced a coefficient of –19.21, indicating a negative association between the two variables. The coefficient of determination (R-squared = 0.188) shows that approximately 19% of the variance in ACT scores is explained by unemployment rate alone. This aligns with the pattern observed in the regression plot and indicates a moderate level of explanatory power.

The second single input model examined the relationship between the percentage of students eligible for free or reduced lunch and average ACT scores. The regression plot displayed a strong negative linear relationship, where schools with higher percentages of students receiving free or reduced lunch tended to have lower ACT scores.

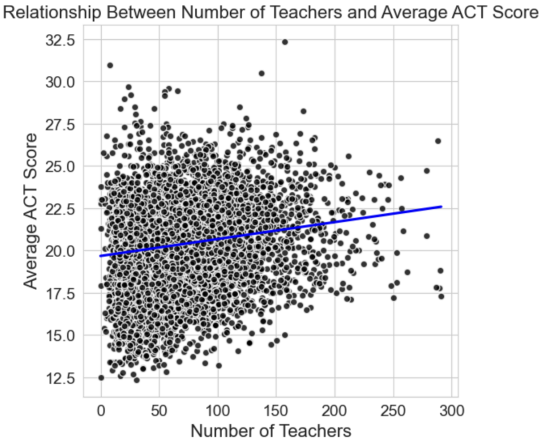


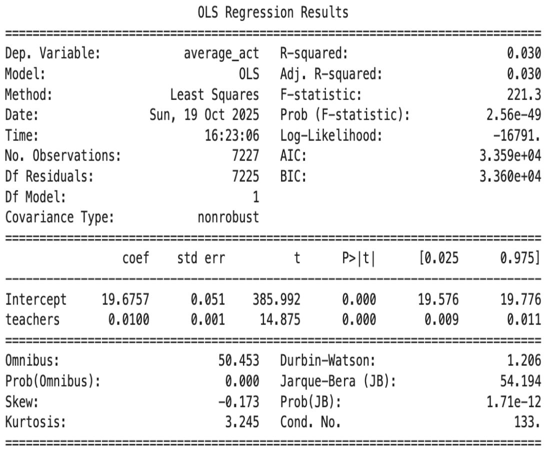


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.614
Model:                            OLS   Adj. R-squared:                  0.614
Method:                 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 fitted simple linear regression model confirmed this association, showing a statistically significant negative relationship (p < 0.001). The estimated coefficient was –0.09, indicating that for each percentage point increase in free or reduced lunch participation, the average ACT score decreased by approximately 0.09 points. The coefficient of determination (R-squared = 0.614) indicates that about 61% of the variance in ACT scores can be explained by this single variable, suggesting a strong model fit consistent with the visual trend. This suggests that the free/reduced lunch percentage is a stronger predictor of ACT performance than the unemployment rate.

The third single-input model examined the relationship between the number of teachers at a school and average ACT scores. The fitted regression model produced a statistically significant positive coefficient (β = 0.01, p < 0.001), indicating a weak positive association between teacher count and ACT performance. The coefficient of determination (R-squared = 0.03) shows that teacher count explains approximately 3% of the variance in ACT scores. The regression plot illustrates a slight upward trend, consistent with the model output, though with substantial scatter around the regression line.





```
                            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

This analysis found that the percentage of students eligible for free or reduced lunch is the strongest predictor of average ACT scores, explaining about 61% of the variation in performance. In comparison, unemployment rate and teacher count were statistically significant but explained much less of the variance. These findings highlight how economic disadvantage at the school level is closely linked to academic achievement.

Several limitations should be considered when interpreting these results. After data preparation, the dataset included information for only about 20 states, which restricts how broadly the findings can be generalized. Additionally, some school-specific factors that likely influence ACT performance—such as tutoring programs, funding levels, and access to academic resources—were not available in the dataset. Having data from multiple school years, rather than just 2016–2017, could also provide a clearer picture of how these relationships change over time.

Despite these constraints, the analysis provides meaningful evidence that school-level economic conditions remain a major determinant of educational outcomes.

Conclusion

This project set out to determine which socioeconomic and school-level factors most strongly 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, explaining over 60% of the variation in ACT performance. This measure directly reflects economic disadvantage at the school level, whereas broader indicators such as unemployment rate or school size (measured by teacher count) were far less informative.

These findings underscore how economic inequality continues to shape academic achievement. If schools, state agencies, and testing organizations aim to improve student success on standardized assessments like the ACT, investing in programs that reduce financial barriers—such as free tutoring, resource access, and preparation tools—could make a meaningful difference. Strengthening school funding and providing additional support for low-income students would help create a more equitable educational environment and allow academic assessments to reflect student ability rather than socioeconomic circumstance.

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