Abstract

This study examines the relationship between student-teacher ratio, socioeconomic factors, and ACT performance among U.S. high schools using data from the National Center for Education Statistics Common Core of Data. While smaller student-to-teacher ratios should theoretically enable more individualized instruction and improve outcomes, our analysis of approximately 7,000 high schools reveals minimal impact. Student-teacher ratio alone explains less than 1% of the variance in ACT scores ($R^2 = 0.002$), despite being statistically significant ($p = 0.001$). In contrast, socioeconomic variables—including unemployment rate, parental higher education, and percentage of students eligible for free or reduced lunch—demonstrated substantially stronger predictive power. When both sets of variables were included in a multiple regression model, student-teacher ratio did not noticeably improve performance suggesting it has limited practical value for predicting ACT scores when socioeconomic factors are considered.

Introduction

Standardized test performance, particularly ACT scores, serves as a critical indicator of college readiness and educational quality in the United States. Education policymakers frequently cite student-teacher ratio as a key lever for improving student outcomes, with the assumption that smaller class sizes enable more individualized instruction and better academic performance (Whitehurst & Chingos, 2011). However, decades of educational research suggest that socioeconomic factors—including household income, parental higher education, and family strcuture —are among the strongest and most consistent predictors of student achievement.

This analysis utilizes data from EdGap.org and National Center for Education Statistics (NCES) Common Core of Data (CCD). All socioeconomic data (household income, unemployment, adult educational attainment, and family structure) included in the EdGap data are from the Census Bureau's American Community Survey. The dataset includes high schools across the United States with the following key variables:

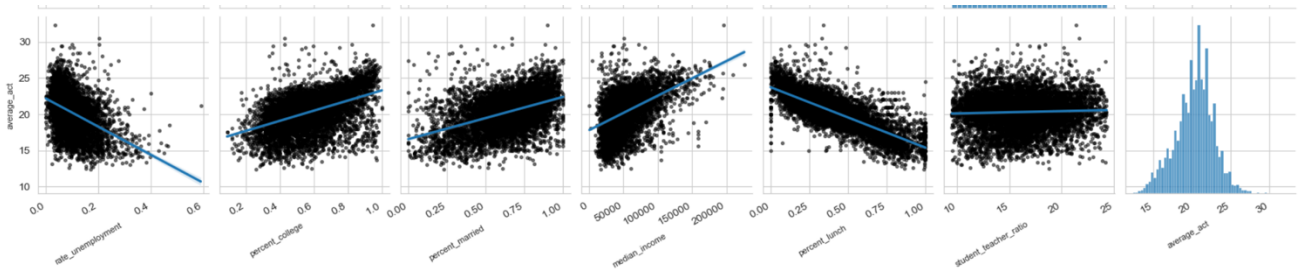
- **NCES School ID:** Unique identifier for each school
- **Student-teacher ratio:** Number of students per instructional staff member
- **Average ACT score:** School-level aggregated ACT score
- **Socioeconomic variables:** Median household income, unemployment rate, percent of adults with college degrees, percent of married adults, and percent of students eligible for free or reduced-price lunch

Data cleaning involved filtering for only high schools and dropping rows with missing ACT data. Missing data for the socioeconomic variables was imputed using an iterative imputer, meaning the missing values were predicted by using the relationships with the other variables. Then outliers were detected and specifically removed in the student-teacher ratio data where only the middle 98% of the data was kept. These values likely represent data entry errors or specialized programs (such as virtual schools) that are not comparable to traditional high schools. It is also important to note that the school data is not evenly spread across the United States, with

a disproportionate number of schools from Texas, which limits the ability to generalize the conclusions of this analysis to nationwide trends.

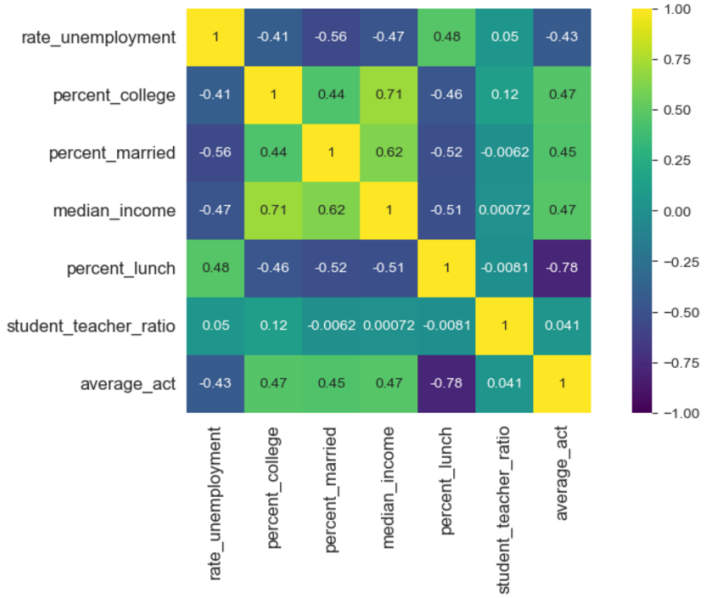
Exploratory Analysis

The exploratory analysis of the data starts with plotting all the variables of interest and their relationships among each other. A regression line is calculated, and the distribution of each variable is displayed so any skewedness or abnormalities can be observed.



The bottom row of the pair plots shows the associations between the socioeconomic variables, student-teacher ratio, and average ACT score. The student-teacher ratio demonstrates weak associations with all variables. There is some noticeable positive correlation between average ACT scores and percent of adults with college degrees, percent of adults who are married, and median income. There is seemingly strong negative correlation between average ACT scores and unemployment rate and percentage of students eligible for free or reduced priced lunch.

The correlation matrix below provides a quantitative view of variable relationships. Strong positive correlations among predictors are evident, particularly between median income and both parental education and married parent percentage. This suggests that not all variables may be necessary useful to the model.

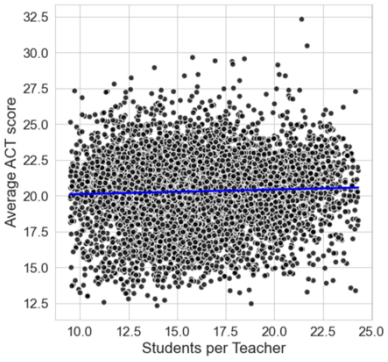


Interestingly, the relationship between student-teacher ratio and ACT score is weakly positive, which is counterintuitive. We might expect lower student-teacher ratios to lead to higher ACT scores, as smaller class sizes theoretically allow teachers to provide more individualized attention. However, this weak positive correlation could be explained by confounding variables. For example, schools in wealthier areas may have higher student-teacher ratios due to larger enrollments while still performing well on the ACT due to access to more resources (e.g. test preparation services) that contribute to student readiness. Additionally, the student-teacher ratio itself may not accurately reflect actual classroom sizes, as it includes all instructional staff.

The weak relationship overall suggests that student-teacher ratio alone is not a strong predictor of ACT performance, and that socioeconomic factors likely play a more prominent role in student achievement.

Modeling

The initial step of the modeling phase was to conduct a simple linear regression with only student-teacher ratio as the predictor and average ACT score as the response variable. This model yielded a R-squared value of 0.002 meaning the predictor, student-teacher ratio, explains 0.2% of the variation in ACT score, thus it has almost no predictive power. The model had a mean absolute error of 1.95 meaning the simple linear regression model can predict ACT scores with an average error of roughly 2 points.



The next step was to construct three multiple linear regression models and compare their accuracy of predicting ACT score. The first (full) model included all predictors: rate of unemployment, unemployment rate, percentage of college graduates, percent of married adults, median income, and percentage of students eligible for free/reduced priced lunch. With this model, it was determined that the percent of married adults and median income variables were not statistically significant which aligns with the relationships seen in the pair plots. Notably, student-teacher ratio was statistically significant ($p < 0.05$), yet as subsequent analysis reveals, this significance does not translate to practical predictive value. This could be due to the large sample size so the relationship is statistically significant but not in a practical sense. The R-squared value of 0.631 indicates 63.1% of the variation in ACT score is explained by the predictor variables. The mean absolute error of 1.13 means the multiple linear regression model with all predictors can predict ACT scores with an average error of roughly 1 point.

The reduced multiple linear regression model, which excludes statistically insignificant variables, yielded identical R-squared and mean absolute error values to the full model. An ANOVA comparison confirmed no statistically significant difference between the models ($p = 0.288 > 0.05$), indicating that the excluded variables contribute no meaningful predictive value.

The final multiple linear regression model excluded the student-teacher ratio variable from the full set of predictors to assess whether it contributed to model accuracy at all. As seen in the table below, the differences in mean absolute error and R-squared between the models are negligible, further reaffirming that student-teacher ratio has virtually no predictive power.

Model	**Mean absolute error**	**R-squared**
Full	1.1315	0.6312
Reduced	1.1314	0.6312
Full excl. student teacher ratio	1.1324	0.6306

Conclusion

In comparing the three models, it is evident that there is minimal difference between the full, reduced, and full excluding student-teacher ratio models. Even though the student-teacher ratio variable is statistically significant, it does not provide any predictive power to the model. In fact, the full model with the student-teacher ratio excluded has the highest mean absolute error, although not by a substantial amount. Therefore, the student-teacher ratio variable does not serve as a good predictor for predicting average ACT score since including it in the multiple regression model did not improve the accuracy.

To address the original question of whether we can predict a school's ACT performance based on socioeconomic factors, the answer is yes. The multiple linear regression models achieve this with an error of approximately one point. Given the range of possible ACT scores, all three models demonstrate a high level of accuracy. However not all variables contribute equally to the model and some variables are not even necessary like student-teacher ratio, median income, and percentage of married adults.

References

EdGap. (n.d.). Educational opportunity and equity data. https://www.edgap.org/

National Center for Education Statistics. (n.d.). Common Core of Data (CCD). U.S. Department of Education. https://nces.ed.gov/ccd/

Whitehurst, G. J., & Chingos, M. M. (2011). Class size: What research says and what it means for state policy. Brookings Institution. https://www.brookings.edu/articles/class-size-what-research-says-and-what-it-means-for-state-policy/