**Is there a relationship between socioeconomic characteristics of school districts and their average ACT scores?**

**Abstract**

This project will create linear regression models between socioeconomic characteristics and average ACT scores in school districts to see if there is a relationship. This is important because ACT scores are one of the main metrics by which student applications to college are judged, and if factors unrelated to the student’s own capabilities are influencing their scores, then it is important to note. This project determines if the socioeconomic characteristics have a relationship with average ACT scores by modeling the linear regression lines with an ordinary least squares model, and checks the p-values, r-squared values, and mean absolute error to determine if there is a statistically significant relationship. It finds that there is a relationship between socioeconomic factors and average ACT scores in school districts. This is important as it demonstrates that factors outside of a student’s control can affect their ACT scores, creating a further disparity between students from good socioeconomic backgrounds to students coming from worse ones.

This project looks to determine if there is a statistically significant relationship between socioeconomic characteristics of school districts and their average ACT scores. To accomplish this task, this project uses data from EdGap.org, which has average ACT scores and five socioeconomic characteristics of school districts across 20 states from the 2016-17 school year. In addition to the five characteristics provided in the EdGap data, this project will also utilize two NCES datasets, one of which includes basic information about the schools, while the other includes an additional characteristic for consideration. The socioeconomic factors considered in this study are:

* Census tract unemployment rate
* Census tract percentage of adults with a college degree
* Census tract percentage of children in a married couple family
* Census tract median household income in dollars
* Percentage of students at the school eligible for free or reduced price lunch
* Ratio of students to teachers at the school.

The three datasets were merged together, and any missing values from the socioeconomic characteristics were imputed using an iterative imputer. This project also filtered out any schools that were not labeled as high schools, as well as removing data that is not impossible (ACT scores above 36 or below 1, percentages less than 0% or greater than 100%, and student-teacher ratios greater than 100). This is later filtered further, removing any data where the student-teacher ratio is greater than 35, in an attempt to eliminate additional outliers. This study is important because ACT scores are essential for applying to and pursuing a college degree, and if socioeconomic characteristics are found to have some sort of statistically significant relationship with average ACT scores, then their use in college applications should be taken into consideration with the circumstances of the students.

To determine whether these socioeconomic characteristic have a statistically significant relationship with ACT scores, this project will be creating 6 linear regression models between each socioeconomic characteristic and the average ACT scores. We are using linear regressions because we are looking at the relationship between each socioeconomic characteristic and the average ACT scores, and linear regression allows us to explain variation in average ACT scores through the variations in the socioeconomic characteristics and allow us to determine whether they have a linear relationship. By fitting these regression lines to an ordinary least squares model, we can check their p-values and use those to determine statistical significance, determining whether there is or is not a relationship between the two variables.

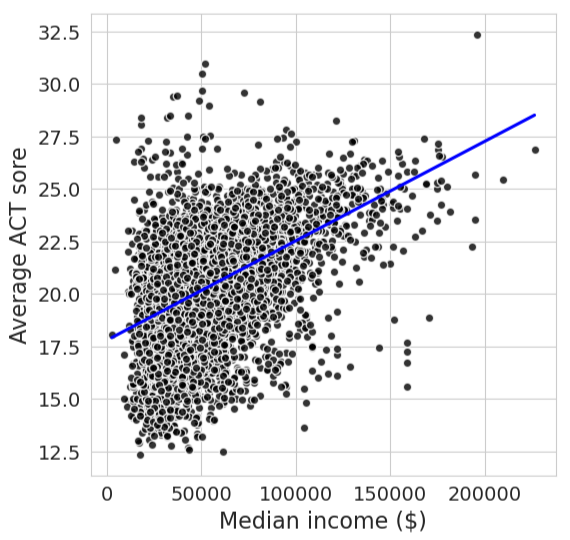
We create linear regression lines for each socioeconomic characteristic, looking at the relationship between the characteristic and the average ACT scores, as shown below:

A graph with a line and a graph

AI-generated content may be incorrect. A graph with a blue line

AI-generated content may be incorrect. A graph with a blue line

AI-generated content may be incorrect. A graph with a line and a blue line

AI-generated content may be incorrect.  A graph with a dotted line

AI-generated content may be incorrect.

While we can see some relationships clearly, such as in the case of percentage of students eligible for free and reduced lunch, other characteristics make it hard to determine whether there is or isn’t a relationship between the two variables, let alone a statistically significant one. To determine statistical significance, this project models the linear regression lines with an ordinary least squares model and then looks towards the p-values to determine whether the relationship between the socioeconomic characteristics and the average ACT scores are statistically significant. In addition to the p-value, we also consider the r-squared value, which is an indicator of how well the model actually relates to the data we’re modeling, and also the mean absolute error, which is a measurement of how far off on average our model is from the actual average ACT scores in the school. These values are presented in the table below:

**Simple Linear Regression Models by Socioeconomic Characteristic**

|  |  |  |  |
| --- | --- | --- | --- |
| **Socioeconomic Characteristic** | **P-Value** | **R-Squared Value** | **Mean Absolute Error** |
| Unemployment Rate | 0.000 | 0.189 | 1.74 |
| Percent College | 0.000 | 0.210 | 1.71 |
| Percent Married | 0.000 | 0.195 | 1.73 |
| Percent Lunch | 0.000 | 0.614 | 1.17 |
| Median Income | 0.000 | 0.213 | 1.71 |
| Student-Teacher Ratio | 0.000 | 0.004 | 1.96 |

Looking at these p-values, we can see that for each individual socioeconomic characteristic, there is a statistically significant relationship between the characteristic and the average ACT scores. While statistically significant, the r-squared values are fairly for all the socioeconomic factors aside from the percentage of students eligible for free and reduced lunch. We can see that most of the variance between the models and the actual data is unaccounted for, especially in the case of the student-teacher ratio, where 99% of the variability in the data is unaccounted for.

The final statistic, mean absolute error, presents how far on average the model is from the actual data. Most of the models are off by around an average ACT score of 1.7 from the actual data, with the percentage of students eligible for free and reduced lunch having a much better mean absolute error, at 1.2, and the student-teacher ratio having a worse mean absolute error, at 2. While there are differences in these mean absolute errors, we can see that even in the worst case, they are within 2 points of the average ACT scores. Considering ACT scores range from 1 to 36, being within 2 points on average and being statistically significant seems to indicate that there is some relationship between the socioeconomic characteristics and the average ACT scores.

Finally, we look to combine all of the socioeconomic characteristics into a single model, and evaluate the overall effect of these characteristics, to try and determine which characteristic is the best predictor for average ACT scores, and whether there is any overlap in correlation between the different characteristics. To accomplish this, we take the six socioeconomic characteristics and create an ordinary least squares model multiple linear regression model, with the statistics shown below:

**Multiple Linear Regression Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Socioeconomic Characteristic** | **P-Value** | **R-Squared Value** | **Mean Absolute Error** |
| Unemployment Rate | 0.000 | 0.629 | 1.14 |
| Percent College | 0.000 | 0.629 | 1.14 |
| Percent Married | 0.000 | 0.629 | 1.14 |
| Percent Lunch | 0.518 | 0.629 | 1.14 |
| Median Income | 0.616 | 0.629 | 1.14 |
| Student-Teacher Ratio | 0.000 | 0.629 | 1.14 |

As we can see in this model, two of our previously statistically significant characteristics are no longer so. This is due to the fact that some of these characteristics are correlated with each other, and so their effect on predicting the average ACT scores is less pronounced when considering all six characteristics. Thus, we remove them from the model, finding that the r-squared values and mean absolute error remain the same. We then graph this model using a residual plot, to evaluate how well the regression model fits the data and see whether the model follows the linear regression as laid out in our model. This graph is shown below:

A graph with black dots

AI-generated content may be incorrect.

As we can see, the data seems to be well represented by our model’s linear regression line, and a relationship between the characteristics and the average ACT scores can be clearly seen.

While we have found evidence of a relationship between the socioeconomic characteristics and the average ACT scores, it is necessary to address one of the main limitations of this project. First, for all but the percentage of students eligible for free and reduced lunch and the student-teacher ratios, the characteristics were using data for the census tracts the school districts were located in, not the schools themselves. This does not account for people living in these census tracts that don’t have students attending school, or who attend schools in other districts. This is probably why the percentage of students eligible for free and reduced lunches seemed to have by far the strongest relationship with average ACT scores. I think despite this, there is still evidence that there is a relationship between socioeconomic factors and average ACT scores, and if anything, the relationship would probably be stronger if we had data from the actual schools themselves rather than the census tracts they reside in.

In conclusion, this project demonstrates a statistically significant relationship between socioeconomic factors and their effects on average ACT scores. This is relevant for any student considering a college degree, as circumstances outside of the control of a student are able to influence and affect their scores in an oftentimes negative way.

References

* EdGap.org. “About the EDGAP Map: Visualizing the Education Gap.” *EdGap.Org | SAT/ACT College Readiness Map*, 2025, www.edgap.org/.
* NCES. “National Center for Education Statistics (NCES).” *National Center for Education Statistics (NCES) | IES,* 2025, nces.ed.gov/