**Exploring Socioeconomic and Regional Factors and Their Influence on   
U.S. High School Educational Outcomes**

Emily Larson

**Abstract**This project explored the impact of socioeconomic and regional predictors on U.S. high schools, using the measurement of average ACT scores and the response variable. A multiple linear regression model was fit using the predictors unemployment rate, percent of adults with college education, median income, parent marriage status, student eligible for free or reduced price lunch, and US region. The analysis discovered that free or reduced price lunch eligibility was the strongest predictor of average ACT score. Other significant predictors were percent adults with college education and unemployment rate. Regional influence was minimal. Median income and parent marriage status were not significant. The results indicate that school-level measurements of socioeconomic predictors are better explanatory variables for predicting average ACT scores than broader area measurements.

**Introduction**This project used data science methodology to explore how socioeconomic factors and regional differences impact educational opportunities in U.S. high schools. Specifically, it used 5 measures of socioeconomic value (unemployment rate, percent adults with college degree, percent children in married couples family, median household income, and percent eligible for free and reduced lunch), with an additional look at regional differences, to determine their value in predicting average ACT scores. The project used socioeconomic data and average ACT scores from EdGap.org and basic school information from the National Center for Education Statistics.

**Theoretical Background**Inequality in educational opportunities for high school students interested in higher education is a common issue in the news and circles of discussion around higher education. Higher education economically advantageous for those who complete degrees. However, with increasingly competitive programs, there are barriers for many students to enter higher education. Students who are at a socioeconomic disadvantage may not have the same access to higher educations and face additional barriers as opposed to students with access to more wealth and resources. One of these barriers is the ACT score, which is a standardized college admissions test. Many colleges choose to include ACT scores in their admissions because it is a measurable indicator of academic performance and a student’s readiness for college.

There is a remaining need to evaluate how a student’s socioeconomic status may influence their access to higher education. One of the ways to do that is to use average ACT scores, a measurable indicator, and evaluate how socioeconomic predictors influence the score. Because a school’s region may impact access to funding and other academic resources, there is also a need to include region in the analysis.

**Methodology**The first data source used was from EdGap, which provided the data for average ACT scores, NCESSCH school id, unemployment rate, percent adults with college degree, percent children in married family, median household income, and percent students that qualify for free or reduced lunch. The second data set, referred to as School Information, was from the National Center for Education Statistics’ Common Core Data for American Public Schools. It was a large dataset containing a multitude of information about school’s identifying information, including school type, NCESSCH school id, and state. Both datasets contained data for the 2016-2017 school year.

Before proceeding with data cleaning and preparation, the basic suitability of the numerical socioeconomic data for the problem was evaluated visually by creating a grid of scatterplots of histograms with regression lines. The graphs demonstrated some observable relationships between the socioeconomic predictors and average ACT scores.

The relevant subsets of data for EdGap and School Information were selected and the columns renamed to easily identify. The two data frames were then joined based on the NCESSCH school id using a left join merge of School Information to EdGap. The left join method was used to ensure all the socioeconomic data from was retained in EdGap. Out of range values, in this case average ACT and percent students that qualify for free or reduced lunch, were removed. Then values in school level were observed and only values for High Schools were retained. At this stage, the two data sets were combined into a single data set. Lastly, it was identified that missing values were present in the dataset. An iterative imputer was used to replace missing values that appears as “NaNs” with values that could be estimated from the dataset.

The number of different states present in the dataset was observed by identifying the unique values and then graphically representing states present on a U.S. map. Dummy variables with values 1 and 0 for different U.S. regions were created based on the region the particular state resides in: Northeast, Midwest, South, and West. An additional column called “region” was created to label the region type of the school.

In the analysis section, each numerical predictor was evaluated individually. A single input regression was fit using STATS model’s ordinary least squares function. The model was then assessed for statistical significance, ability to predict average ACT scores, and error of the model. There was focus on the magnitude of each coefficient. Each numerical predictor was also assessed if it needed to be transformed. This was done by plotting the residuals and visually examining to see if they form a perfect cloud shape or other discernable shape. For the regional variables, a boxplot was generated to visually observe the range of average ACT scores per region.

In the case of unemployment rate, possible heteroscedasticity was observed when viewing the residuals. A logarithmic (“log”) transformation was performed on the variable for unemployment rate and then the model assessed using similar methods used on the initial model. After comparing the regression summary and mean absolute error for both the original model and log transformed model, it was determined that the log-transformed model did not perform better than the original model in predicting average ACT scores.

A multiple linear regression model was then fit using all the numerical predictor and dummy variables. The model was evaluated by looking at the R-squared value, magnitude of the coefficients, and statistical significance of each predictor. A reduced model that omitted the insignificant predictors median income and perfect of parents married was then created. The mean absolute error and R-squared of the reduced model and full model were then compared to determine if the omitted variables meaningfully improved the full model.

Lastly, a model was fit using standardized predictors based on the predictors utilized in the reduced models. The numerical predictors were standardized to have a mean of 0 and standard deviation of 1. This was done to compare the relative importance of each predictor in the reduced model. The dummy variables for region were not included in the standardization process because they already present as values 0 and 1. A multiple linear regression model was then fit using the normalized predictor and regional dummy variables. The fit and accuracy of the reduced model and standardized model was compared by using the mean absolute error and R-squared.

**Computational Results**Table 1: Summary of full multiple linear regression model  
Note: R-squared = 0.63 | Mean Absolute Error= 1.14

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **p-value** |
| Intercept | 22.72 | 0.00 |
| Unemployment Rate | -2.25 | 0.00 |
| Percent Adults with College Degree | 1.65 | 0.00 |
| Percent Children in Married Family | -0.14 | 0.29 |
| Median Household Income | 0 | 0.28 |
| Percent Eligible for Free or Reduced Lunch | -7.58 | 0.00 |
| Northeast Region | -1.18 | 0.00 |
| West Region | 0.33 | 0.00 |

Table 2: Summary of reduced multiple linear regression mode  
Note: R-squared = 0.63 | Mean Absolute Error= 1.14

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **p-value** |
| Intercept | 22.65 | 0.00 |
| Unemployment Rate | -2.17 | 0.00 |
| Percent Adults with College Degree | 1.73 | 0.00 |
| Percent Eligible for Free or Reduced Lunch | -7.58 | 0.00 |
| Northeast Region | -1.18 | 0.00 |
| West Region | 0.33 | 0.00 |

Figure 1: Residual Plot of Full Regression Model

A graph with black dots

AI-generated content may be incorrect.

Figure 2: Residual Plot of Reduced Model

A graph with black dots

AI-generated content may be incorrect.

**Discussion**The purpose of this project was to examine whether socioeconomic factors and regional differences influence educational opportunities in U.S. high schools, using average ACT scores as a measurable indicator of educational outcomes. Predictor variables included both numerical and categorical variables, such as unemployment rate, percent of adults with college degree, percent of students with married parents, median income, percent of students eligible for free or reduced-price lunch, and region. Both the response variable and predictors were measurable and contained a sufficient number of measurements for analysis.

One limitation in the data was the scope of geographic representation. The dataset included schools from only 20 states, which somewhat restricted the ability for a complete analysis of all US regions. Many rural and less densely populated states were excluded, limiting the model's ability to evaluate different types of socioeconomic locations. This is particularly relevant because rural poverty may differ significantly from urban poverty, potentially biasing the results.

The analysis produced quantitative results that helped evaluate the relationship between the predictors and average ACT scores. The model showed that unemployment rate, percent of adults with college degree, and percent of students eligible for free or reduced lunch were statistically significant predictors. Regional variables were also significant but had a relatively small effect. Median income and percent of students with married parents were not significant contributors to the model.

The percent of students eligible for free or reduced-price lunch was the strongest predictor of average ACT scores. In contrast, median income was the weakest. This is likely because free and reduced lunch eligibility is a direct, school-level measure of economic disadvantage within schools. Median income is an area-level measurement that may include households without school-aged children and does not account for income distribution or cost of living. Therefore, free or reduced lunch percentage better captures the socioeconomic context directly affecting students, making it a stronger predictor in the model.

**Conclusion**  
The project demonstrated that socioeconomic factors, especially the percentage of students eligible for free or reduced price lunch, are strong predictors of the average ACT score in U.S. high schools. The region of the high school has a small influence. While other factors like parental education and unemployment rate also showed significant relationships, variables such as median income and marital status of parents did not meaningfully improve the model. This suggests that school level measurements of student socioeconomic wellbeing are more effective at explaining academic outcomes than broader measurements.

**References**Memphis Teacher Residency. (2016). *EdGap: SAT/ACT college‑readiness map*. Retrieved October 19, 2025, from <https://www.edgap.org/#5/37.718/-95.998>

National Center for Education Statistics. (School Year 2016-2017). *Common Core of Data (CCD): Public Elementary/Secondary School Universe Survey data*. U.S. Department of Education. Retrieved October 19, 2025, from [https://nces.ed.gov/ccd/pubschuniv.asp](https://nces.ed.gov/ccd/pubschuniv.asp?utm_source=chatgpt.com)