Socioeconomic Impact on ACT Standardized Test Scores

**Abstract:**

We analyzed the impact of six socioeconomic factors on the results of ACT scores across the US. We measured unemployment rate, married family percentage, Percentage of parents with a college education or higher, percentage of schools with free or reduced lunch programs, median income, and a school’s Title I eligibility status.

Using exploratory data analysis methods and statistical regression models, we found that schools with free or reduced lunch programs had the strongest relationship of the six variables in their impact on ACT scores. Additionally, percent married, median income, and Title I were not statistically significant when tested in a multiple linear regression model.

**Introduction:**

Socioeconomic factors significantly impact everyone. The median income in the area where you live or the rate of unemployment in a region can have a significant impact on everyone’s quality of life. Education and access to higher learning can have a profound impact on the trajectory of an individual’s life. Standardized testing in the US is often used as a step in determining which colleges or universities will accept a student for admission. However, this raises the question of whether socioeconomic variables can influence someone’s standardized test scores. ***For this study, we aim to apply data science methodology to investigate whether socioeconomic factors impact standardized test results in the United States.***

**Data Collection & Cleaning:**

In this case, we are going to compare average ACT scores in school districts to the socioeconomic factors of unemployment rate, the percentage of parents obtaining a college degree, the percentage of married couples, the median income of an area, the percentage of free or reduced school lunch programs, and Title I eligible schools from the year 2016-2017.

We pulled data from EdGap.org, an organization that tracks average high school SAT and ACT scores, median household income, and other socioeconomic indicators (edgap.org), as well as the National Center for Education Statistics (NCES). We cleaned the data by merging our EdGap datasets with our NCEI data using an NCESSCH ID as a connecting variable between our three data frames. We cleaned our N/A data using an iterative imputer and converted our categorical Title I data to a numerical value to filter out N/A data in the imputer. Some ACT data required adjustments to its parameters because the minimum and maximum ACT scores fell below the range of 0 to 36.

**Theoretical Background:**

To answer the question of whether socioeconomic factors impact ACT scores, we plan to utilize a combination of descriptive and diagnostic approaches. At a descriptive level, pair plots and scatter plots will be especially useful because they allow us to examine large datasets at a high level while simultaneously tracking correlation data within these data frames.

Regarding our diagnostic approach, we will test hypotheses in regression models to help us make inferences about relationships within our data. This will involve us creating predictive models and testing them against our actual results to analyze statistical significance.

Regression models such as polynomial linear regression, reduced models, and multiple linear regression will help us analyze our factors individually and in tandem with other variables and our ACT scores. In these models, we will pay specific attention to our R-squared data, coefficient data, and P-value to help us answer our question.

**Methodology (after import and initial analysis:**

Using a comprehensive pair plot that included all our variables helped us understand the key relationships between our variables. It answered the question of whether this data was worth analyzing in greater detail.

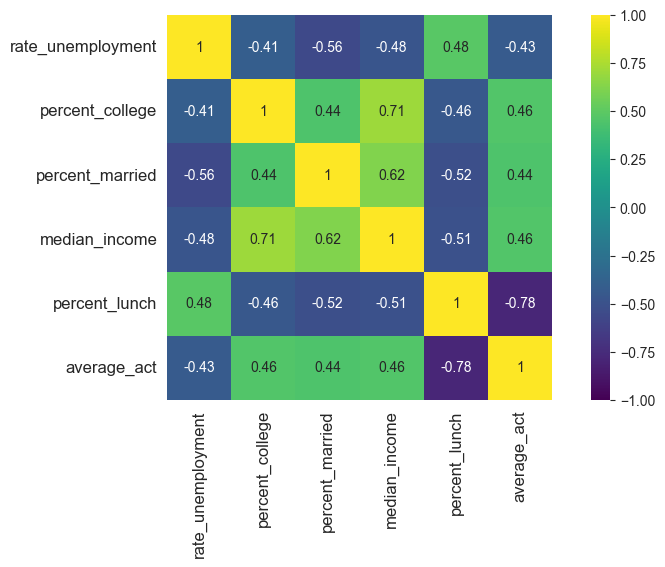
Quality control involved ensuring that our data fell within the correct parameters, which included analyzing the details of our categorical variables, such as state data. Lastly, we imputed any missing data in our data frames and exported a clean dataset for our initial exploratory analysis.

For our exploratory data analysis, we used a heatmap and a comprehensive pair plot to visualize our key numerical predictor variables in the initial correlation data.

Our modeling consisted of running a single-input model first, then pivoting to our multiple linear regression model, followed by a reduced model based on statistically significant socioeconomic factors, and finally a scaled reduced model to gain a better understanding of our coefficients.

**Computational Results:**

Our heatmap showed an initial correlation analysis of our key numerical factors:



These are the results of our multiple linear regression analysis:

OLS Regression Results

==============================================================================

Dep. Variable: average\_act R-squared: 0.628

Model: OLS Adj. R-squared: 0.628

Method: Least Squares F-statistic: 2032.

Date: Wed, 22 Oct 2025 Prob (F-statistic): 0.00

Time: 20:32:30 Log-Likelihood: -13327.

No. Observations: 7227 AIC: 2.667e+04

Df Residuals: 7220 BIC: 2.672e+04

Df Model: 6

Covariance Type: nonrobust

=====================================================================================

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------

Intercept 22.7267 0.141 161.664 0.000 22.451 23.002

rate\_unemployment -2.2667 0.404 -5.609 0.000 -3.059 -1.474

percent\_college 1.7368 0.158 11.026 0.000 1.428 2.046

percent\_married -0.0774 0.134 -0.579 0.563 -0.340 0.185

median\_income -1.073e-07 1.21e-06 -0.089 0.929 -2.48e-06 2.26e-06

percent\_lunch -7.6055 0.097 -78.549 0.000 -7.795 -7.416

Title\_1 -0.0396 0.039 -1.019 0.308 -0.116 0.037

==============================================================================

These are the results of the mean absolute errors of the reduced regression model compared to our full multiple linear regression model:

|  | **Mean Absolute Error** | **R-squared** |
| --- | --- | --- |
| **full model** | 1.1453 | 0.6280 |
| **reduced model** | 1.1615 | 0.6185 |

And here are the results of our scaled reduced model:

OLS Regression Results

==============================================================================

Dep. Variable: average\_act R-squared: 0.628

Model: OLS Adj. R-squared: 0.628

Method: Least Squares F-statistic: 4063.

Date: Wed, 22 Oct 2025 Prob (F-statistic): 0.00

Time: 19:14:03 Log-Likelihood: -13328.

No. Observations: 7227 AIC: 2.666e+04

Df Residuals: 7223 BIC: 2.669e+04

Df Model: 3

Covariance Type: nonrobust

================================================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------------------------

Intercept 20.2986 0.018 1127.575 0.000 20.263 20.334

rate\_unemployment\_normalized -0.1227 0.021 -5.800 0.000 -0.164 -0.081

percent\_college\_normalized 0.2826 0.021 13.503 0.000 0.242 0.324

percent\_lunch\_normalized -1.7770 0.022 -81.992 0.000 -1.819 -1.735

==============================================================================

Omnibus: 873.370 Durbin-Watson: 1.483

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3131.099

Skew: 0.587 Prob(JB): 0.00

Kurtosis: 6.004 Cond. No. 1.93

==============================================================================

|  | **Mean Absolute Error** | **R-squared** |
| --- | --- | --- |
| **normalized model** | 1.1455 | 0.6279 |
| **reduced model** | 1.1615 | 0.6185 |

**Discussion:**

Initially, our heatmap revealed the strongest correlation between our percentage of free/reduced lunch programs and ACT scores, with a correlation coefficient of -0.78. Our other four variables showed moderate correlations across the board. For our multiple linear regression model, we found that Title I, percent married, and median income did not show statistically significant results.

The reduced model posted an R-squared value of 0.6185 and a mean absolute error of 1.1615, which was not a significant difference compared to the multiple linear regression results (0.6279 and 1.1455). The scaled model produced intriguing results, as we found that in a scaled model where the mean of the factors is 0 and the standard deviation is equal to 1, a change in the factor of free/reduced lunch correlated with a 1.177-point reduction in ACT scores. This relationship was the strongest of the three coefficients measured.

**Conclusion:**

Regarding our initial question of whether socioeconomic variables have an impact on ACT scores, we found statistically significant relationships with the variables of the percentage of parents who attended college, unemployment rate, and the percentage of schools with free lunch, with the latter showing the strongest correlation among all the variables. Additionally, we found no statistically significant relationship between a school’s Title I eligibility and its ACT scores. The percentage of Married Couples and Median income also did not show statistically significant results in our multiple linear regression model. While these analyses demonstrate statistical significance, it is worth noting that other factors can also impact an individual's performance on a standardized test. However, this report provides a clearer indication of the relationships between these socioeconomic factors and test performance.

**References:**

Secondary Information System, Elementary and. “Elementary and Secondary Information System.” *Elsi - Elementary and Secondary Information System*, National Center for Education Statistics, 2002, nces.ed.gov/ccd/elsi/.

Teacher’s Residency, Memphis. “About the EDGAP Map: Visualizing the Education Gap.” *EdGap.Org | SAT/ACT College Readiness Map*, https://memphistr.org/, 1 Jan. 2016, www.edgap.org/#5/37.718/-95.998.