ABSTRACT

This study examines which socioeconomic factors are most closely associated with average ACT scores in 2016–2017. A state-level dataset was built by left-joining EdGap with two NCES datasets (school directory and state finance), with data available for 20 states. Rows missing ACT values were removed, and missing values for socioeconomic predictors were imputed using multivariate iterative imputation. The analysis included exploratory summaries, a single-predictor income model, and multiple linear regressions, including a reduced model with standardized predictors. The income-only model explained little variance ($R^2 = 0.219$). With all numerical predictors included, fit improved to $R^2 = 0.628$. In the reduced model, the share of students receiving free or reduced-price lunch had the largest (negative) coefficient, and state total revenue had the smallest (negative) coefficient. Overall, economic hardship factor was the most informative correlates of cross-state differences in ACT performance.

INTRODUCTION

Differences in average ACT scores across states reflect more than classroom instruction. Household stability and material resources shape the environments in which students learn and test. This project asks a clear scientific question: which state-level socioeconomic indicators are most strongly associated with ACT performance in 2016–2017?

Data come from EdGap and the National Center for Education Statistics (NCES). EdGap provides school IDs and several socioeconomic indicators. Two NCES data sets are used: a school directory (identifiers and categories) and a state finance file (we use state total K–12 revenue). The EdGap table is the primary dataset; the NCES school directory is left-joined to EdGap on school ID, and the NCES finance table is left-joined on state. Left joins ensure all EdGap records are retained while adding matching attributes from NCES. Data come from EdGap and the National Center for Education Statistics (NCES). EdGap provides school IDs and several socioeconomic indicators. Two NCES datasets are used: a school directory (identifiers and categories) and a state finance file (using state total K–12 revenue). EdGap is the primary table. The NCES school directory is left-joined to EdGap on school ID, and the NCES finance table is left-joined on state, so all EdGap records are retained while matching attributes from NCES are added. The analytic sample includes 20 states that are available. Missing values are imputed only for the socioeconomic predictors using a multivariate iterative imputation.

THEORETICAL BACKGROUND

Higher income is expected to be positively related to ACT performance because it often brings access to supports that aid learning, such as stable housing, tutoring, reliable internet, and time. Higher unemployment can create stress and reduce resources, which likely lowers performance. Family background and structure also matter: higher marriage rates can indicate greater household stability and supervision, and higher adult college-degree rates reflect communities where education is common and adults can assist with schoolwork. The number of students receiving free or reduced-price lunch indicates economic disadvantage. When this number is higher, students typically have fewer outside-of-school supports and scores tend to be lower. Finally, total state revenue indicates the overall resources available to schools and could be positively related to performance, though its effect depends on how funds are allocated and local costs.

METHODOLOGY

The analysis begins with exploratory data analysis to check whether socioeconomic factors relate to average ACT scores and to confirm that the data are suitable for modeling. Columns relevant to the study are then selected and renamed for clarity. All datasets are combined with the EdGap table using left joins so that every EdGap record is retained. Specifically, the NCES school directory is joined to EdGap on school ID, and the NCES finance table is joined on state. Unreasonable or out-of-range values are set to NaN, and the dataset is restricted to high schools. Predictor columns with gaps are completed using a multivariate iterative imputation imputation that estimates each predictor from the others. State total revenue is then updated by dividing each state's total funding by the number of high schools in that state to obtain a rough per-school estimate.

Next, a correlation matrix and pair plot of the numerical variables are produced to explore relationships among variables. Modeling starts with single-input models: a simple linear regression is fit first, followed by a quadratic version to see whether accuracy improves. A multiple linear regression using all socioeconomic variables is then fit, and model fit and accuracy are assessed. A reduced model containing only the significant predictors is fit last. Finally, all predictors are normalized so that coefficients are comparable across variables.

RESULTS

Figure 1. Correlation matrix

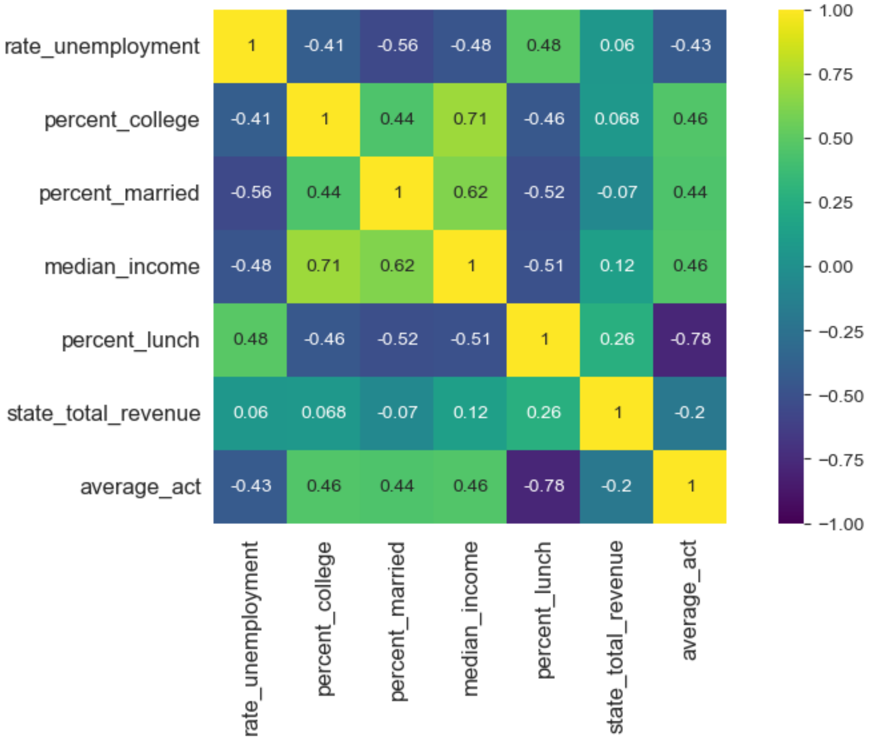


Figure 1 shows the correlation heatmap reports pairwise Pearson coefficients. With average_act, the coefficients are: percent_lunch r=-0.78; median_income r=0.46; percent_college r=0.44; percent_married r=0.46; rate_unemployment r=-0.43; and state_total_revenue r=-0.20.

Figure 2. Average ACT score vs. median household income

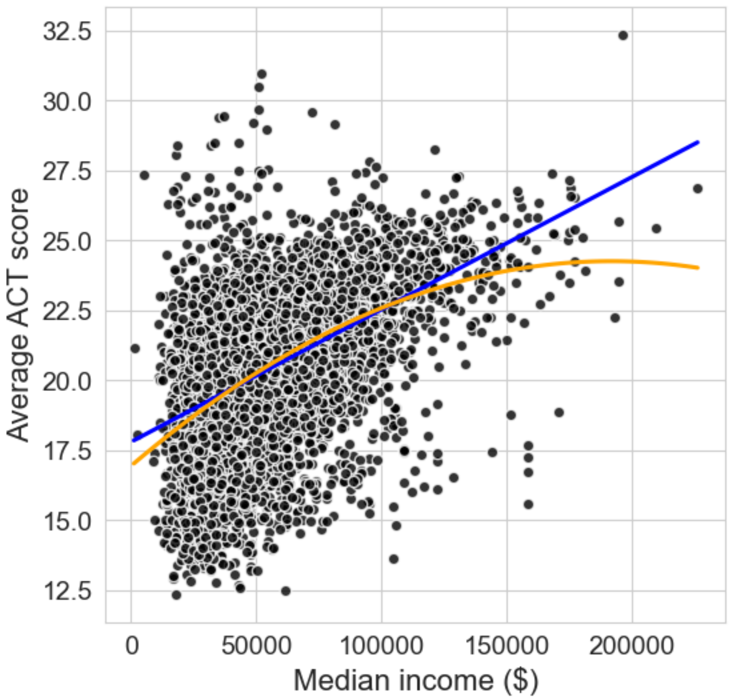
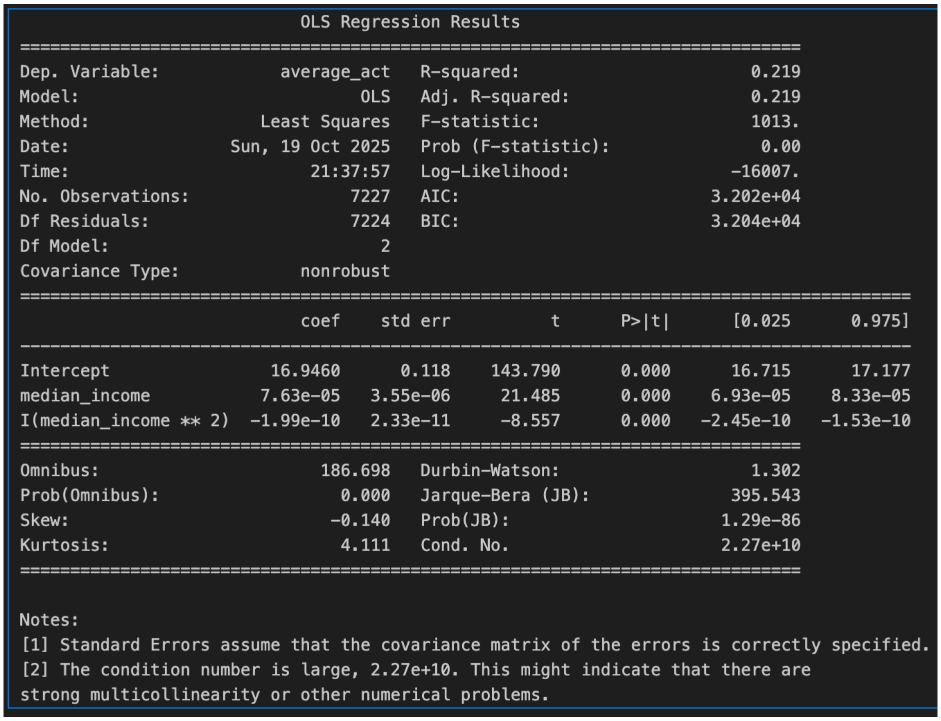


Figure 2 show a scatter plot of average ACT score against median household income. The straight blue line is the simple OLS fit ACT~income .The orange curve is a quadratic fit ACT~income^2.

Table 1. Single-predictor model (ACT ~ median income)

Table 1

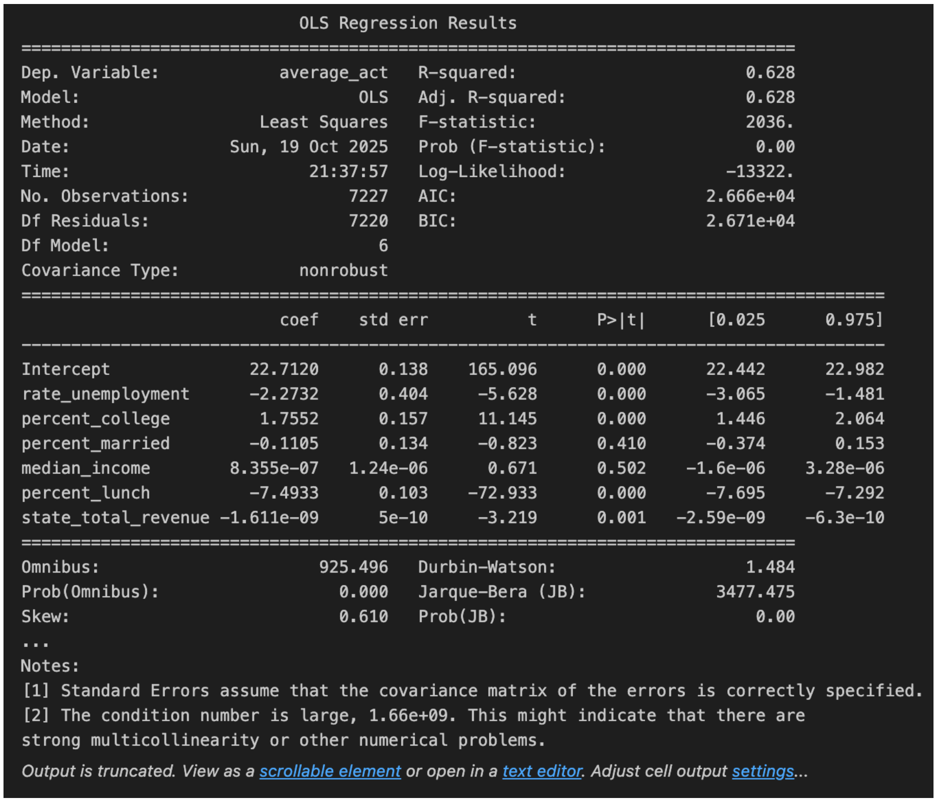


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.219
Model:                            OLS   Adj. R-squared:                  0.219
Method:                 Least Squares   F-statistic:                     1013.
Date:                Sun, 19 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:37:57   Log-Likelihood:                -16007.
No. Observations:                7227   AIC:                         3.202e+04
Df Residuals:                    7224   BIC:                         3.204e+04
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept                16.9460      0.118    143.790      0.000      16.715      17.177
median_income           7.63e-05   3.55e-06     21.485      0.000    6.93e-05    8.33e-05
I(median_income ** 2) -1.99e-10   2.33e-11     -8.557      0.000   -2.45e-10   -1.53e-10
==============================================================================
Omnibus:                      186.698   Durbin-Watson:                   1.302
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              395.543
Skew:                          -0.140   Prob(JB):                     1.29e-86
Kurtosis:                       4.111   Cond. No.                     2.27e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.27e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 1 shows a quadradic model of average ACT score on median household income yields R^2 = 0.211.

Table 2. Multiple linear regression model with all numerical predictors

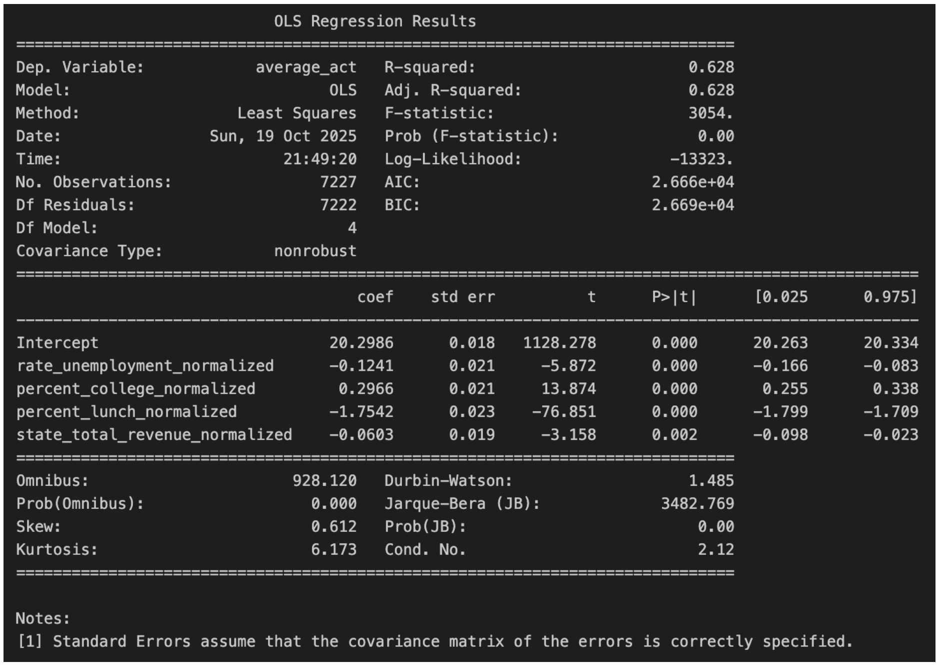


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.628
Model:                            OLS   Adj. R-squared:                  0.628
Method:                 Least Squares   F-statistic:                     2036.
Date:                Sun, 19 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:37:57   Log-Likelihood:                -13322.
No. Observations:                7227   AIC:                         2.666e+04
Df Residuals:                    7220   BIC:                         2.671e+04
Df Model:                           6
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept              22.7120      0.138    165.096      0.000      22.442      22.982
rate_unemployment      -2.2732      0.404     -5.628      0.000      -3.065      -1.481
percent_college         1.7552      0.157     11.145      0.000       1.446       2.064
percent_married        -0.1105      0.134     -0.823      0.410      -0.374       0.153
median_income        8.355e-07   1.24e-06      0.671      0.502    -1.6e-06    3.28e-06
percent_lunch          -7.4933      0.103    -72.933      0.000      -7.695      -7.292
state_total_revenue -1.611e-09      5e-10     -3.219      0.001   -2.59e-09    -6.3e-10
==============================================================================
Omnibus:                      925.496   Durbin-Watson:                   1.484
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3477.475
Skew:                           0.610   Prob(JB):                         0.00
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Table 2 shows that the multiple linear regression model yeilds R^2 = 0.628. Predictor p-values are less than 0.001 for all variables except percent_married (p = 0.410) and median_income (p = 0.502).

Table 3. Reduced multiple linear regression model with normalized predictors



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.628
Model:                            OLS   Adj. R-squared:                  0.628
Method:                 Least Squares   F-statistic:                     3054.
Date:                Sun, 19 Oct 2025   Prob (F-statistic):               0.00
Time:                        21:49:20   Log-Likelihood:                -13323.
No. Observations:                7227   AIC:                         2.666e+04
Df Residuals:                    7222   BIC:                         2.669e+04
Df Model:                           4
Covariance Type:            nonrobust
====================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
Intercept                         20.2986      0.018   1128.278      0.000      20.263      20.334
rate_unemployment_normalized      -0.1241      0.021     -5.872      0.000      -0.166      -0.083
percent_college_normalized         0.2966      0.021     13.874      0.000       0.255       0.338
percent_lunch_normalized          -1.7542      0.023    -76.851      0.000      -1.799      -1.709
state_total_revenue_normalized    -0.0603      0.019     -3.158      0.002      -0.098      -0.023
==============================================================================
Omnibus:                      928.120   Durbin-Watson:                   1.485
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3482.769
Skew:                           0.612   Prob(JB):                         0.00
Kurtosis:                       6.173   Cond. No.                         2.12
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 3 reports the reduced multiple linear regression with normalized predictors, yielding R^2 = 0.628.

DISSUSION

The income-only model shows an upward pattern but low explanatory power (R^2=0.219). Including all numerical predictors improves fit to R^2=0.628. In the full model, the share of students receiving free or reduced-price lunch has the largest negative coefficient, indicating that economic hardship aligns with lower average scores; the share of adults with a college degree is positive and sizable, reflecting community educational level associated with higher scores; and unemployment is negative but smaller, consistent with short-term economic stress playing a secondary role. Median income and marriage rate contribute little once the other variables are included ($p>0.05$), suggesting overlap with other stronger indicators in the model. State total revenue has the smallest and negative coefficient, indicating that statewide totals doesn't show how funds are distributed and used.

CONCLUSION

This study set out to identify which socioeconomic factors are most strongly associated with state average ACT scores in 2016–2017. The models indicate that the shares of students receiving free or reduced-price lunch and adults with a college degree have the highest correlation to average ACT scores across states. These findings suggest that addressing achievement gaps requires attention to the local burden of economic hardship and the educational context in which students live, not just overall funding levels.

REFRENCES

[1] EdGap. (2017). EdGap socioeconomic indicators, 2016–2017 [Data set]. EdGap. (File: EdGap_data.xlsx)

[2] U.S. Department of Education, National Center for Education Statistics. (2017). Common Core of Data (CCD): Public elementary/secondary school universe survey, 2016–2017 [Data set]. (File: ccd_sch_029_1617_w_1a_11212017.csv)

[3] U.S. Department of Education, National Center for Education Statistics. (2017). State education finance data, 2016–2017 [Data set]. (File: Stfis170_1a.xlsx)