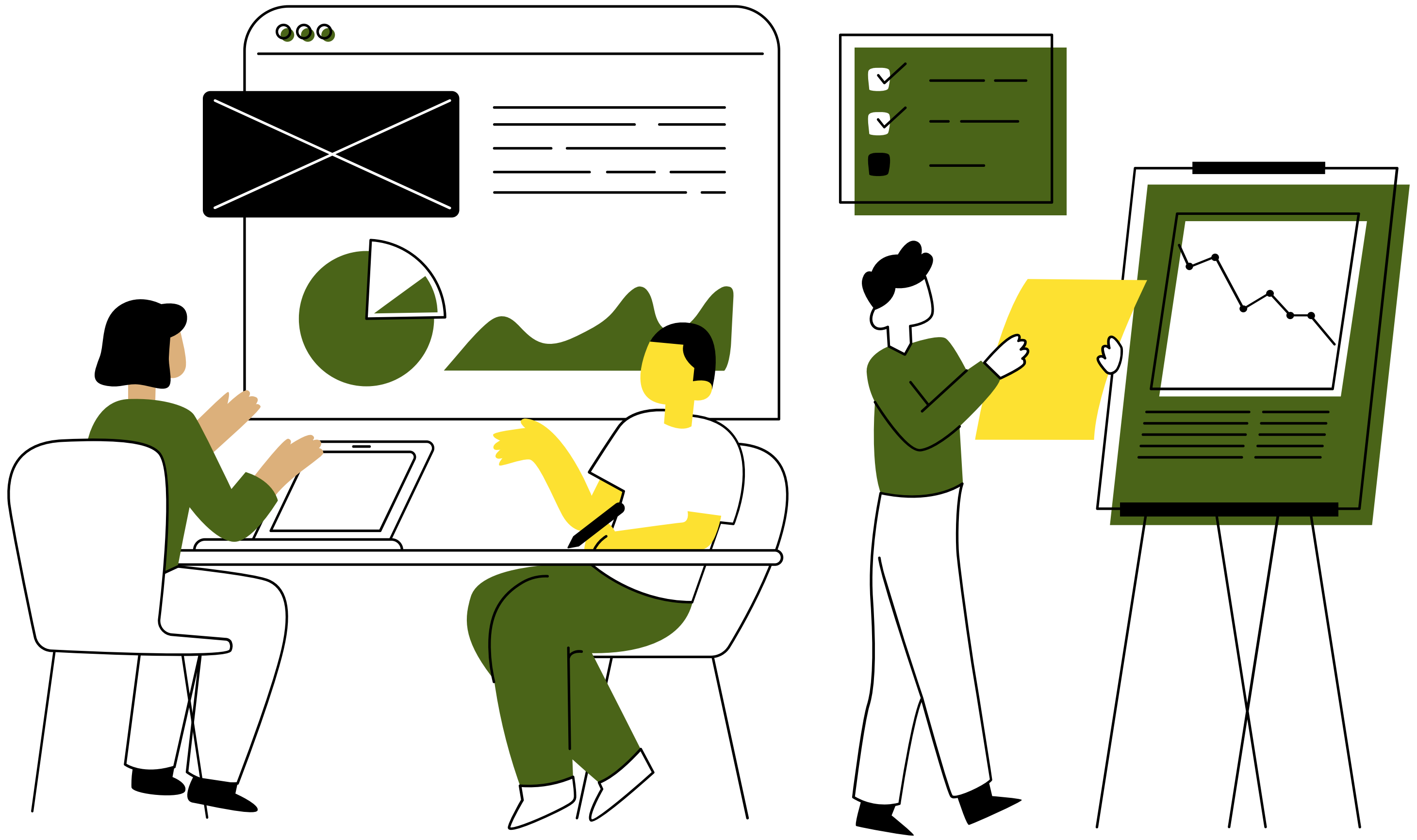
Socioeconomic Factors vs. ACT Performance

DATA 3320 - Tina Chau

Introduction

- ACT and SAT are standardized exams taken by high school students to apply to college
- Aiming to explore correlations between socioeconomic factors and performance on the ACT
- Dataset includes several socioeconomic factors, such as median income, college attendance rate, free/reduced lunch rate, unemployment rate, and marriage rate.



Description of the Data

- National Center for Education statistics
    - Includes 20 states, with 7,227 schools
    - Data collected from 2016-2017
- Edgap.org
    - ACT and SAT score averages across states collected from 2016-2017
- EdWeek.org
    - Provides information on what states require SAT/ACT scores for college in 2017

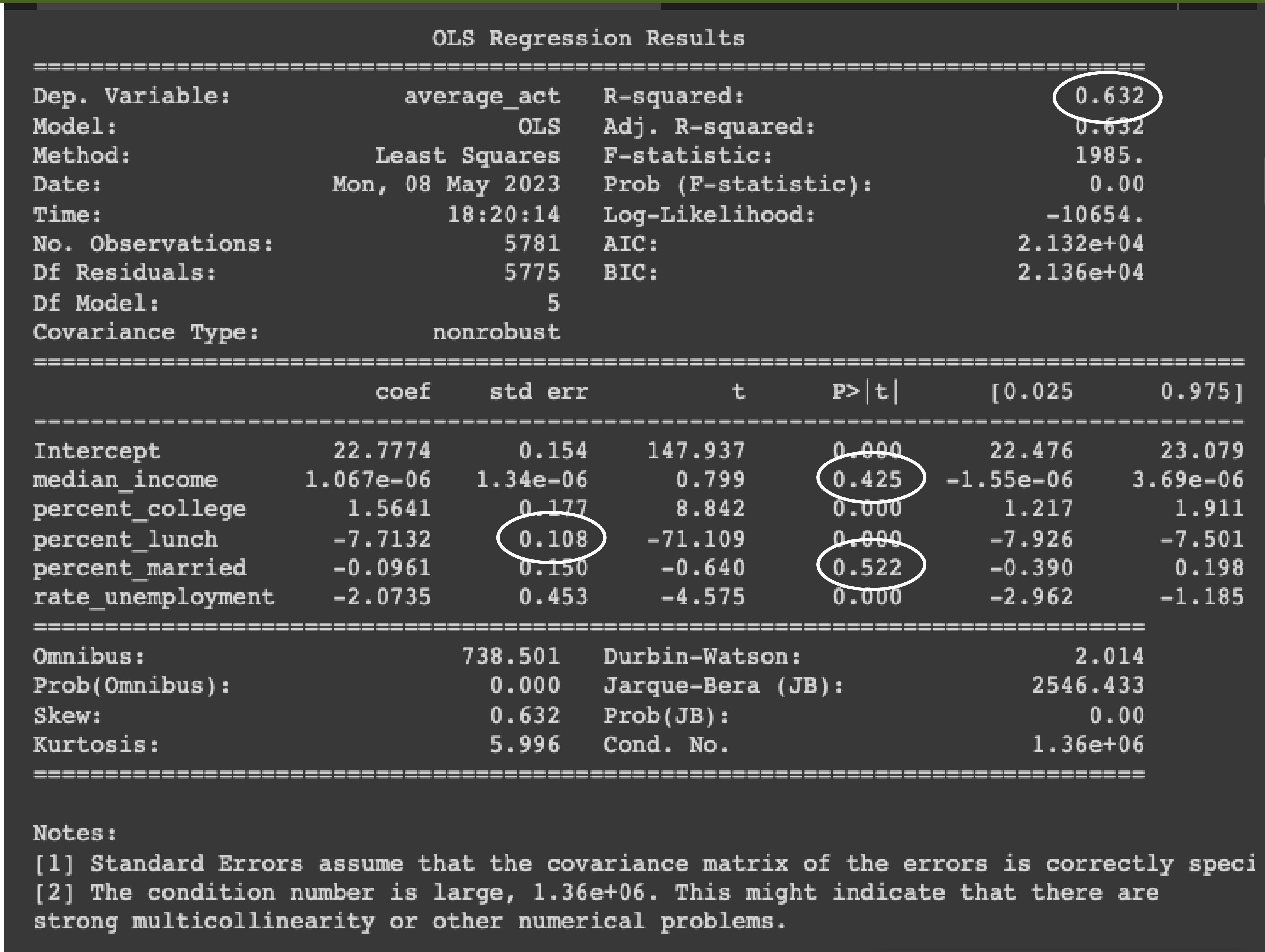
Data Science Questions

- Is there a correlation between socioeconomic factors and student's ACT score performances in their area?
    - What predictors have the strongest correlation to average_act?
    - And do the variables indicate statistical significance?

Methods of Analysis

Multiple Linear Regression

- High r-squared value, indicates good fit
- Lowest standard of error is percent_lunch with 0.108
- Based on p-value, median_income and percent_married are not statistically significant



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.632
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     1985.
Date:                Mon, 08 May 2023   Prob (F-statistic):               0.00
Time:                        18:20:14   Log-Likelihood:                -10654.
No. Observations:                5781   AIC:                         2.132e+04
Df Residuals:                    5775   BIC:                         2.136e+04
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            22.7774      0.154    147.937      0.000      22.476      23.079
median_income      1.067e-06   1.34e-06      0.799      0.425   -1.55e-06    3.69e-06
percent_college       1.5641      0.177      8.842      0.000       1.217       1.911
percent_lunch        -7.7132      0.108    -71.109      0.000      -7.926      -7.501
percent_married      -0.0961      0.150     -0.640      0.522      -0.390       0.198
rate_unemployment    -2.0735      0.453     -4.575      0.000      -2.962      -1.185
==============================================================================
Omnibus:                      738.501   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2546.433
Skew:                           0.632   Prob(JB):                         0.00
Kurtosis:                       5.996   Cond. No.                     1.36e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly speci
[2] The condition number is large, 1.36e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

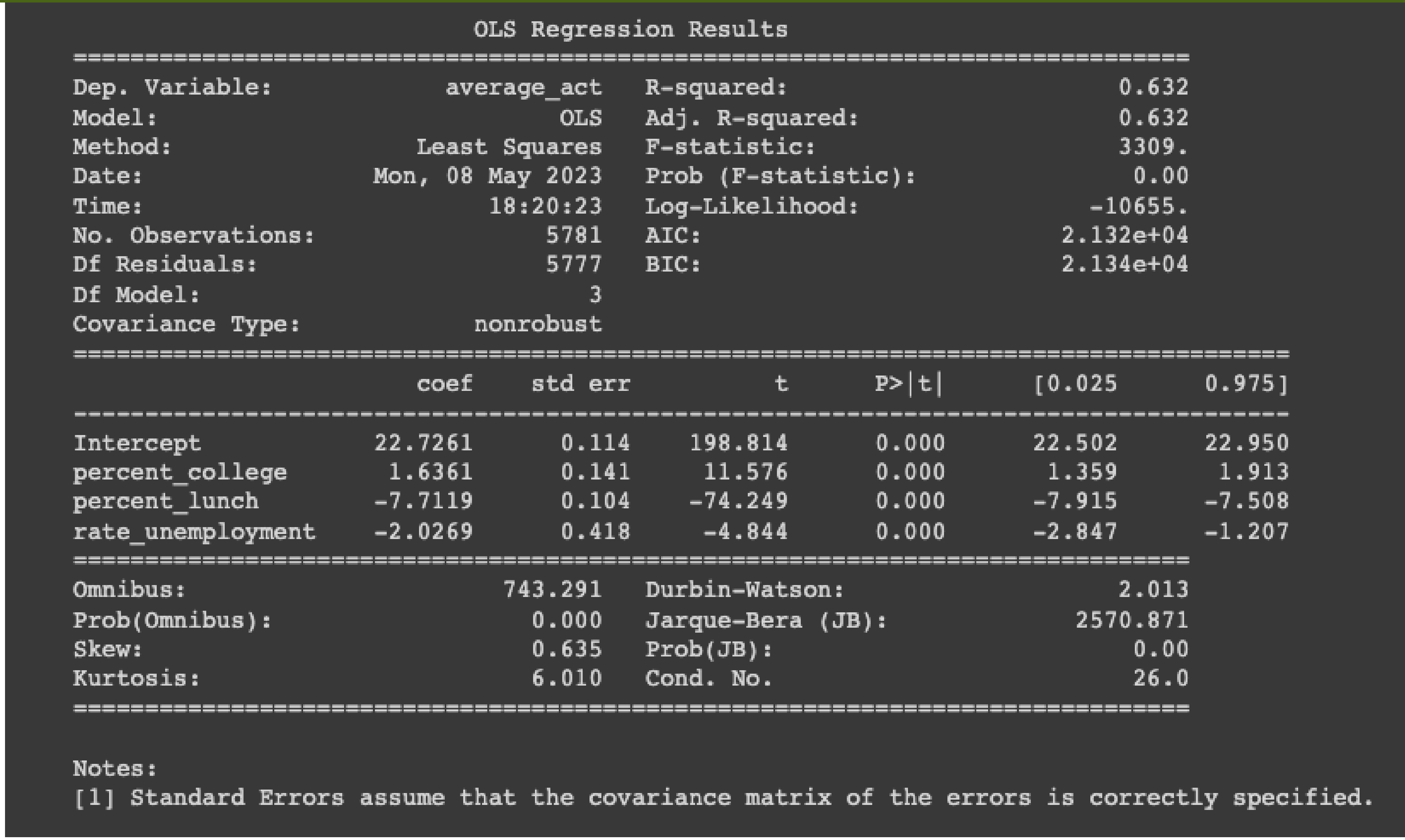
Best Subset Selection

The best combination:

- percent_college
- percent_lunch
- rate_unemployment
- on average_act

Why?

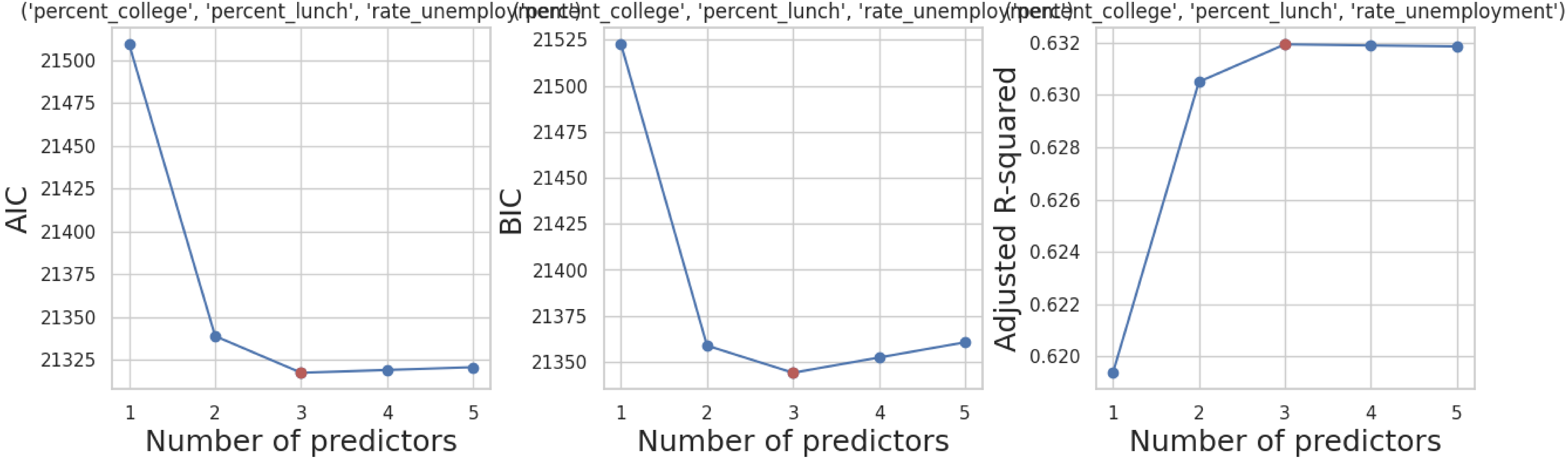
- High r-squared value of 0.632
- P-value below level of significance



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.632
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     3309.
Date:                Mon, 08 May 2023   Prob (F-statistic):               0.00
Time:                        18:20:23   Log-Likelihood:                -10655.
No. Observations:                5781   AIC:                         2.132e+04
Df Residuals:                    5777   BIC:                         2.134e+04
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            22.7261      0.114    198.814      0.000      22.502      22.950
percent_college       1.6361      0.141     11.576      0.000       1.359       1.913
percent_lunch        -7.7119      0.104    -74.249      0.000      -7.915      -7.508
rate_unemployment    -2.0269      0.418     -4.844      0.000      -2.847      -1.207
==============================================================================
Omnibus:                      743.291   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2570.871
Skew:                           0.635   Prob(JB):                         0.00
Kurtosis:                       6.010   Cond. No.                         26.0
==============================================================================

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

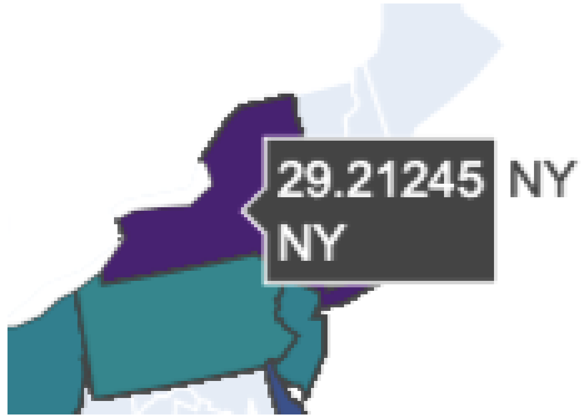
Best Subset Selection

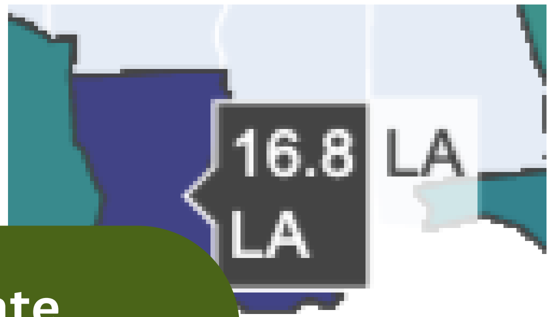


- best predictors highlighted in red
- important to note that lower values of AIC and BIC indicate a better fit
- a higher value of r-squared indicates a stronger predictive performance.

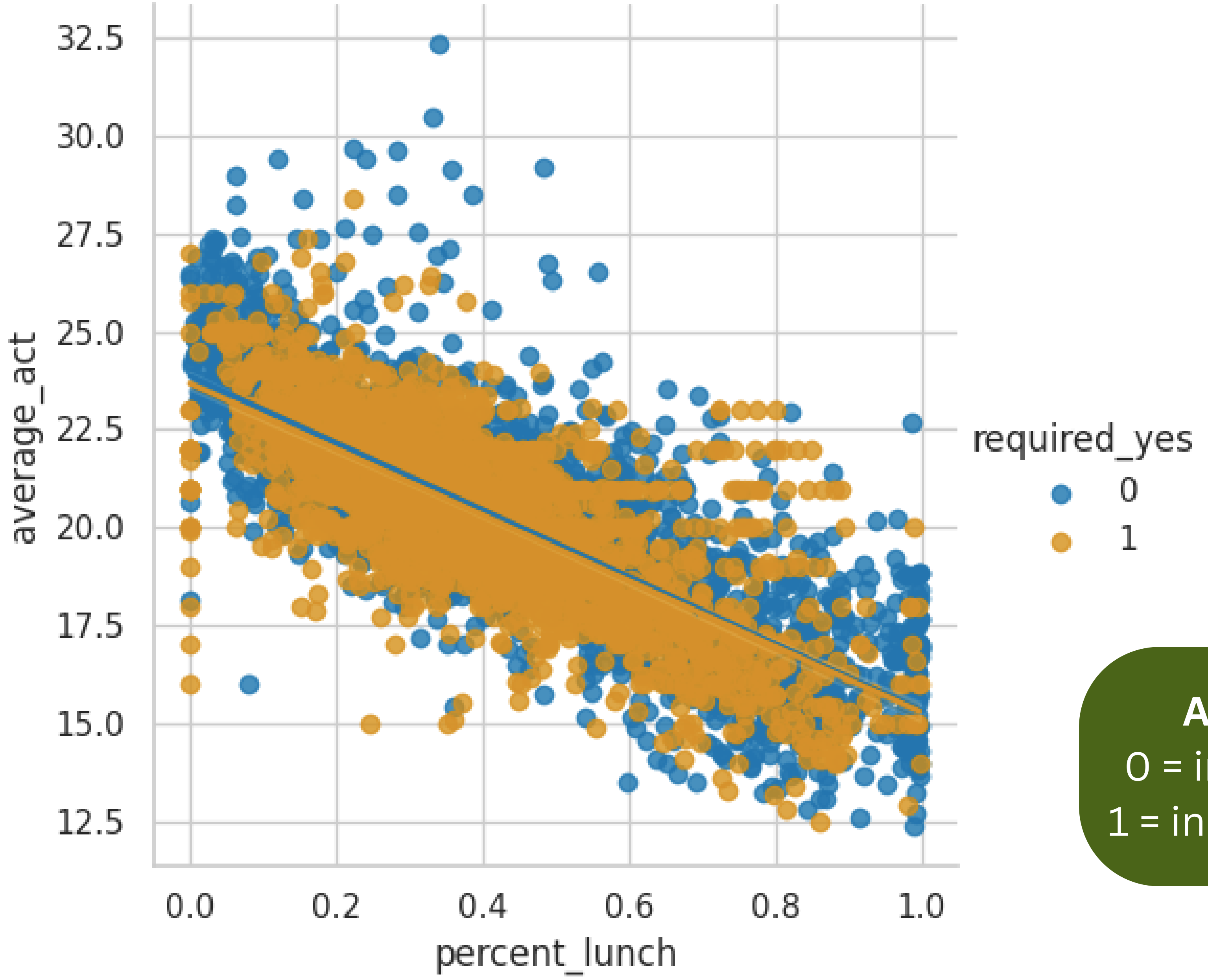
Additional Step to the Project

- What is the relationship between the requirement for students to take the ACT/SAT and the average scores on these tests across different states?





ACT Average by State
New York: highest average, 29.2.
Alabama: lowest average, 16.8

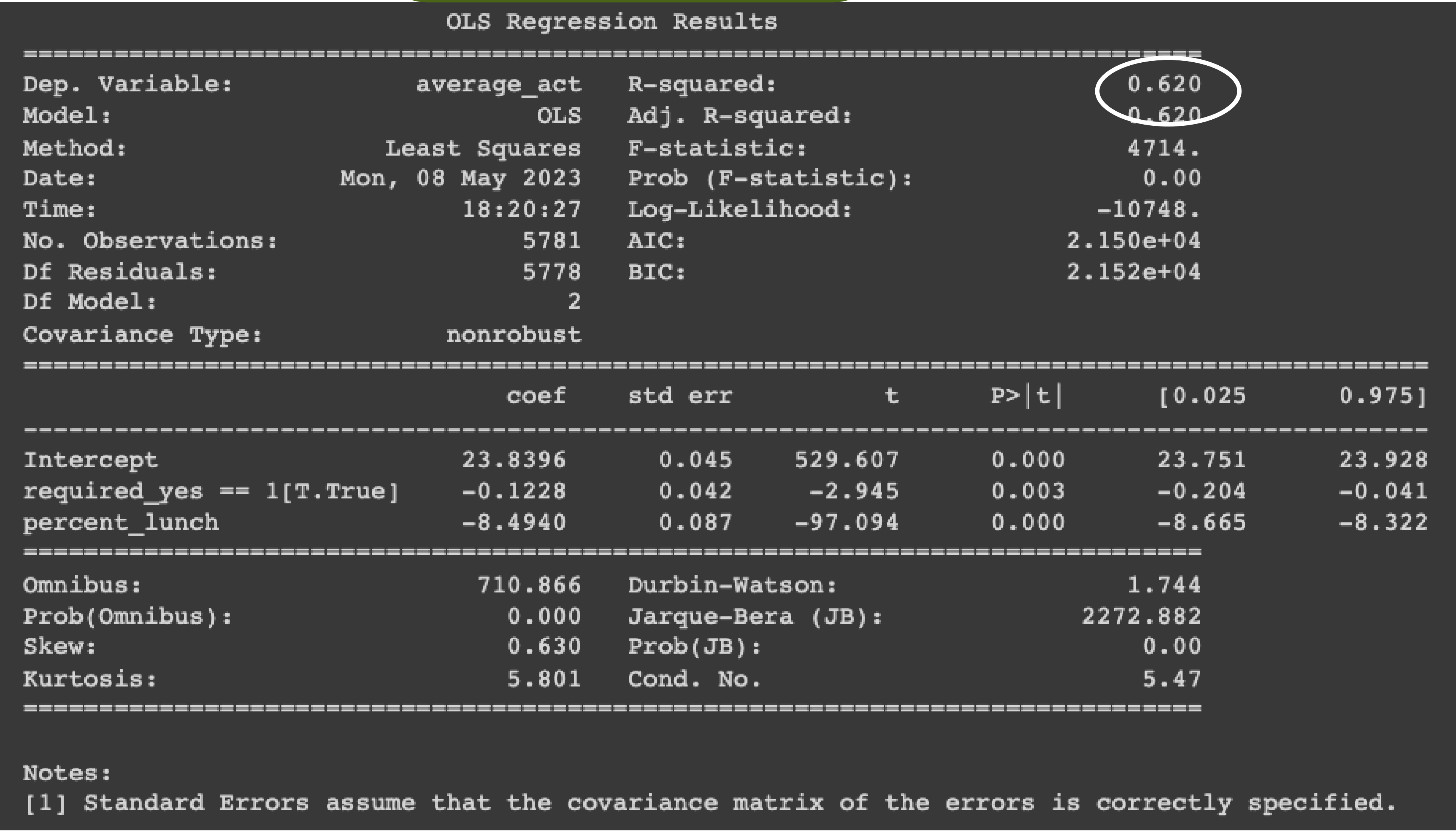


ACT Requirement by State
0 = indicates states don't require
1 = indicates states that do require

Additional Step to the Project

- Since some state have requirement and some don't, does this affect the relationship between ACT performance and socioeconomic factors? Can the inclusion of ACT/SAT requirements help counteract this bias?

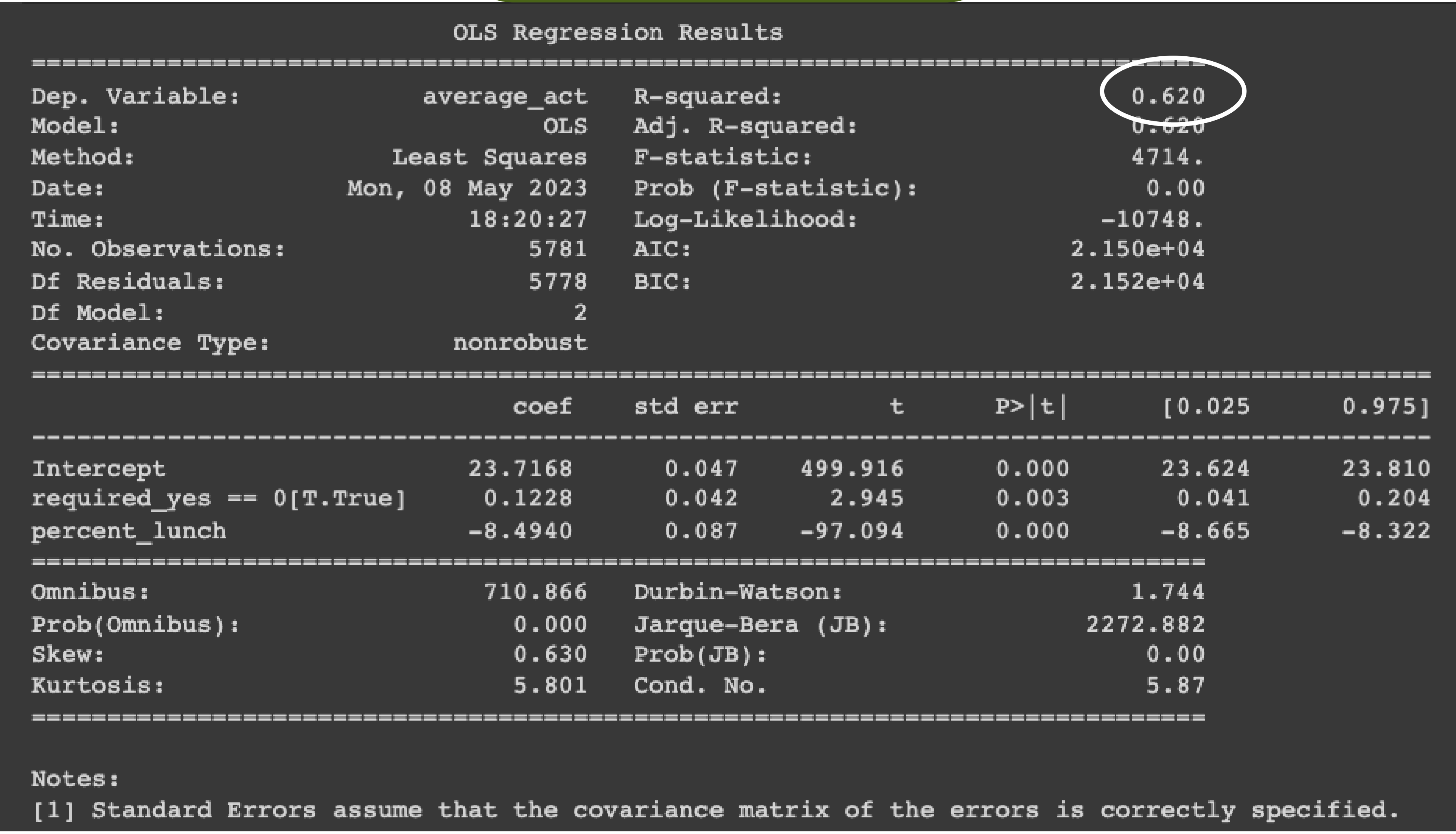
Multiple Linear Regression
Includes states that require



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.620
Method:                 Least Squares   F-statistic:                     4714.
Date:                Mon, 08 May 2023   Prob (F-statistic):               0.00
Time:                        18:20:27   Log-Likelihood:                -10748.
No. Observations:                5781   AIC:                         2.150e+04
Df Residuals:                    5778   BIC:                         2.152e+04
Df Model:                           2
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                    23.8396      0.045    529.607      0.000      23.751      23.928
required_yes == 1[T.True]    -0.1228      0.042     -2.945      0.003      -0.204      -0.041
percent_lunch                -8.4940      0.087    -97.094      0.000      -8.665      -8.322
=============================================================================================
Omnibus:                      710.866   Durbin-Watson:                   1.744
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2272.882
Skew:                           0.630   Prob(JB):                         0.00
Kurtosis:                       5.801   Cond. No.                         5.47
==============================================================================

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

Multiple Linear Regression
Includes states that don'trequire



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            average_act   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.620
Method:                 Least Squares   F-statistic:                     4714.
Date:                Mon, 08 May 2023   Prob (F-statistic):               0.00
Time:                        18:20:27   Log-Likelihood:                -10748.
No. Observations:                5781   AIC:                         2.150e+04
Df Residuals:                    5778   BIC:                         2.152e+04
Df Model:                           2
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                    23.7168      0.047    499.916      0.000      23.624      23.810
required_yes == 0[T.True]     0.1228      0.042      2.945      0.003       0.041       0.204
percent_lunch                -8.4940      0.087    -97.094      0.000      -8.665      -8.322
==============================================================================
Omnibus:                      710.866   Durbin-Watson:                   1.744
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2272.882
Skew:                           0.630   Prob(JB):                         0.00
Kurtosis:                       5.801   Cond. No.                         5.87
==============================================================================

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

Additional Step to the Project

- Since some state have requirement and some don't, does this affect the relationship between ACT performance and socioeconomic factors? Can the inclusion of ACT/SAT requirements help counteract this bias?
    - The difference in r-squared value isn't much
    - P-value is below the level of significance 0.05, which does indicate statistical significance.
    - However, we can conclude that ACT requirements do not impact the relationship between ACT performance and socioeconomic factors by much.

Conclusion

- We found that ACT performance is heavily influenced by socioeconomic factors such as percent_lunch, percent_college, and rate_unemployment.
- After analyzing both states that require and do not require ACT scores, we found that there is a negative relationship between average_act and percent_lunch. The regression results did not show any significant difference in r-squared values between the two groups, and the difference was only 0.001.
- Out of all the states in the dataset, New York had the highest average ACT score of 29.2, while Alabama had the lowest score of 16.8. Despite missing some states in the dataset, we found that there is a strong relationship between percent_lunch and average_act with a high r-squared value.