Education Project

Abstract

The goal of this project was to discover what socioeconomic factors predict academic success as measured by average ACT score. The data used includes the average ACT score of the school, unemployment rate, percentage of adults with a college degree, percentage of children in a married couple family, median household income in dollars, and percentage of the students at the school eligible for free or reduced lunch, and the number of homeless students in the district.

Regression analysis was used to discover the socioeconomic variables that predict average ACT score. The socioeconomic factors of unemployment rate, percentage of adults with a college degree, and percentage of students at the school eligible for free or reduced lunch together predict academic success by predicting the average ACT score of the students at the school.

Introduction

Is school performance, measured by ACT score, predicted by socioeconomic factors? This is the question asked in this project. The goal of this project is to discover what socioeconomic factors predict average ACT score.

The primary data used comes from EdGap.org, an organization that maps average ACT scores and socioeconomic factors at high schools. This data includes ACT scores (or SAT scores converted to an equivalent ACT score), unemployment rate, percentage of adults with a college degree, percentage of children in a married couple family, median household income in dollars, and percentage of the students at the school eligible for free or reduced lunch. Although it originally came from EdGap.org, I downloaded it from GitHub repository for DATA 5100.

The secondary data is from the National Center for Education Statistics at https://nces.ed.gov/ccd/pubschuniv.asp. This data includes the school year, the state, the school name, the school ID number, the school district ID number, the zip code, the type of school, and the school level (such as high school or elementary school). Although it originally came from the National Center for Education Statistics, I downloaded it from a Dropbox link.

The tertiary data is from Data Express and can be downloaded at https://eddataexpress.ed.gov/download. This data includes the number of homeless students in the school district and the school district ID. Data can be downloaded from many different years. I downloaded the data from the 2016-2017 school year since that was the year my EdGap data and school information data was from.

The EdGap data and the school information data were merged on the school ID. This combined dataframe was then merged with the homeless student data on the school district ID. Negative percentages and ACT scores were changed to null values. Rows with missing average ACT scored were deleted. An iterative imputer was used to replace missing values in the socioeconomic data. The exception to this was the homeless data. So much of the homeless

data was missing that a separate data frame was created for evaluating how the homeless data predicted the average ACT score.

This study is important because all children deserve the opportunity to succeed. When some schools are not producing students capable of achieving an ACT score that shows their readiness for college, students are failed by the school system. If socioeconomic factors that predict ACT scores can be identified, states and school districts can use that information to close the learning gap between schools. The aim of this project is to find the socioeconomic factors that predict student achievement, measured in this case by ACT scores.

Theoretical Background

Regression analysis will be used to determine the socioeconomic factors that predict average ACT scores. The relationship between socioeconomic factors and ACT scores is what this project seeks to understand. Regression analysis allows an equation to be created that describes the relationship between the dependent variable, average ACT scores, and the independent variables, the socioeconomic factors.

Methodology

First I used Seaborn to plot a scatter plot and regression line of median income versus average ACT score. I used statsmodels ordinary least squares function with the formula version to fit the simple linear regression. I displayed the fit summary and used that to examine the coefficient on median income. I examined the p-value and R-squared. I calculated the root mean square error and the mean absolute error using metrics from scikit-learn. I used a residual plot for graphical analysis of model fit. Because there was some structure to the plot, I tried a quadratic model. I plotted the regression curves and the scatter plot using Seaborn. I fit a quadratic linear regression model. I displayed the fit summary and assessed the model significance. Then I used an analysis of variance (ANOVA) to compare the simpler model to the more complicated model and determine whether the difference between the more complicated model model and the simpler model is statistically significant. I used anova_lm from statsmodels to do this. I also found the mean absolute error of the quadratic model.

At this point I went on to fit a multiple linear regression model using the socioeconomic variables as predictors. I used statsmodels ordinary least squares function with the formula version to fit the multiple linear regression and printed the summary. I used a residual plot for graphical assessment of model fit. I computed the mean absolute error. After finding that some of the predictors were not statistically significant, I fit a reduced model with the significant predictors: unemployment rate, percentage of adults with a college degree, and percentage of the students at the school eligible for free or reduced lunch. I used a residual plot for graphical assessment of the model fit. I performed a numerical assessment of the accuracy using mean absolute error. I compared the accuracy between the full and reduced models using mean absolute error.

Then I scaled the predictor variables in the reduced model so I could use the magnitude of the coefficients in this model to compare the relative importance of each of those predictor variables at contributing to the estimate of the average ACT score. I used StandardScalar from scikit-learn to find the correct transformation to apply. I then applied that scalar to the predictor values. Then

I fit the multiple linear regression model with the normalized predictors. I looked at the fit summary. I compared the mean absolute error and the R-squared of the normalized model and the reduced model.

I found that the percent of student eligible for free or reduced lunch was the strongest predictor in the model in terms of how much a change in that variable is contributing to an estimated change in the ACT score. Because of this, I decided to make a single input model for percent free and reduced lunch. I plotted the regression line and the scatter plot using Seaborn. I fit the simple linear regression using statsmodels ordinary least squares function with the formula version. I displayed the fit summary. I performed a numerical fit summary of the fit accuracy by finding the root mean square error and the mean absolute error. I used a residual plot for graphical analysis of the model fit.

I then plotted a regression line and scatter plot of the homeless data versus the average ACT score. I fit the simple linear regression using statsmodels ordinary least squares function with the formula version. I displayed the fit summary.

Computational Results

The R-squared of the simple linear regression of median income as a predictor of average ACT score was 0.211. The statistical significance of the coefficient on our predictor, the p-value, was 0.000. The root mean square error (RMSE) was 2.228. The mean absolute error was 1.713.

The R-squared of the quadradic linear regression model of median income as a predictor of average ACT score was 0.219. The coefficient on the squared term was statistically significant, with a p-value of 0.000. The mean absolute value was 1.697. The results of the ANOVA are in Figure 1.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	7225.0	35865.012794	0.0	NaN	NaN	NaN
1	7224.0	35505.105960	1.0	359.906834	73.227974	1.395848e-17

Figure 1 ANOVA comparing the two nested polynomial linear regression models.

The R-squared of the simple linear regression of percent of students eligible for free or reduced lunch as a predictor of average ACT score was 0.614. The statistical significance of the coefficient, the p-value was 0.000. The root mean squared error (RMSE) was 1.559. The mean absolute error was 1.169.

The R-squared of the simple linear regression of the number of homeless students as a predictor of average ACT score was 0.059. The statistical significance of the coefficient, the p-value was 0.000.

The R-squared of the multiple linear regression of the unemployment rate, percentage of adults with a college degree, percentage of children in a married couple family, median household income in dollars, and percentage of the students at the school eligible for free or reduced lunch as predictors for average ACT score was 0.628. The statistical significance of the coefficient for unemployment rate, the p-value, was 0.000. The statistical significance of the coefficient for

percentage of adults with a college degree, the p-value, was 0.000. The statistical significance of the coefficient for percentage of children in a married couple family, the p-value, was 0.577. The statistical significance of the coefficient for median household income, the p-value, was 0.920. The statistical significance of the coefficient for percentage of the students at the school eligible for free or reduced lunch, the p-value, was 0.000. The mean absolute error for this model was 1.145.

The R-squared of the reduced model, the multiple linear regression of the unemployment rate, percentage of adults with a college degree, and percentage of the students at the school eligible for free or reduced lunch as predictors for average ACT score was 0.628. All three variables had p-values of 0.000. The mean absolute error for this model was 1.145. A comparison of the mean absolute error and the R-squared for the full model and the reduced model are presented in Figure 2.

	Mean Absolute Error	R-squared
full model	1.1453	0.6280
reduced model	1.1455	0.6279

Figure 2 Comparison of the Mean Absolute Error and the R-squared for the full model and the reduced model.

The R-squared of the scaled model was 0.628. The coefficient for the normalized unemployment rate was -0.1227. The coefficient for the normalized percentage of adults with a college degree was 0.2826. The coefficient for the normalized percentage of the students at the school eligible for free or reduced lunch was -1.7770.

Discussion

Based on the simple linear regression, it appeared that median income could predict average ACT score to some extent based on the R-squared of 0.211, the statistical significance of the coefficient on our predictor, the p-value, of 0.000, the root mean square error (RMSE) of 2.228, and the mean absolute error of 1.713.

After fitting a quadratic linear regression model, it appeared that a linear model would be sufficient to predict the ACT score, considering that a quadratic was not necessarily going to provide a much better fit. This was based on the similar mean absolute errors.

The multiple linear regression of the unemployment rate, percentage of adults with a college degree, percentage of children in a married couple family, median household income in dollars, and percentage of the students at the school eligible for free or reduced lunch as predictors for average ACT score had a relatively high R-squared of 0.628. However, two of the model coefficients were not statistically significant, based on their relatively large p-values. Thus, another multiple linear regression was created with these two variables removed. The R-squared of this reduced multiple linear regression was the same, 0.628.

After scaling, there was a much larger magnitude of a coefficient for the percent lunch variable than for either of the other two variables. This says that the estimated change in the average

ACT score at a school is of much larger magnitude when we would have a one standard deviation change in the percent lunch variable as compared to the percent college or the unemployment rate variable.

Because of this, I created a single input model to examine whether the percentage of the students at the school eligible for free or reduced lunch was a better predictor for average ACT score by itself than as part of a multiple linear regression. The percentage of the students at the school eligible for free or reduced lunch was a strong predictor of average ACT score. The model had an R-squared of 0.614. However, this R-squared was not as high as the R-squared of the reduced multiple linear regression.

Before I came to a conclusion, I wanted to examine how well the number of homeless students in the school district predicted the average ACT score. The number of homeless students in the school district did not predict the average ACT score very well, as the model had an R-squared of 0.059.

The reduced multiple regression showed that the unemployment rate, percentage of adults with a college degree, and percentage of the students at the school eligible for free or reduced lunch together were the best predictors of average ACT score, based on the R-squared of 0.628.

Conclusions

The socioeconomic factors of unemployment rate, percentage of adults with a college degree, and percentage of students at the school eligible for free or reduced lunch together predict academic success by predicting the average ACT score of the students at the school. This is important information for the school districts, the state and local governments, and educational advocacy groups to have. They can use this information as a basis for doing more projects that dive deeper and wider into the data to better understand what changes will help lower performing schools bridge the gap between them and higher performing schools.

References

Ed Data Express. (2025). *SY1617_FS195_DG814_LEA.csv* [Data set]. U.S. Department of Education. https://eddataexpress.ed.gov/download.

Fischer, Brian. (2025). [“Average ACT or SAT scores for schools and several socioeconomic characteristics of the school district downloaded from EdGap.org”] [Data set]. Seattle University, College of Science and Engineering.

Fischer, Brian. (2025). [“School information data from the National Center for Education Statistics”] [Data set]. Seattle University, College of Science and Engineering.