Abstract

The Education Project seeks to identify which socioeconomic and educational variables are the strongest predictors of average ACT scores. Exploratory data analysis and statistical modeling were applied to a compilation of US school data on ACT scores, school type, and socioeconomic predictors. Results showed that the percentage of students on free or reduced-price lunch was the most significant predictor of average ACT scores at a given school. These results could outline the basis for prescriptive analysis on improving ACT scores, but further research may be necessary.

Introduction

The Education Project is a data analysis project designed to assess how strongly ACT scores of US students are correlated with a variety of socioeconomic factors, the type of school the students are attending (e.g. private, public, charter, etc.), and whether their school receives Title I assistance from the federal government. This analysis required the compiling of three datasets, each containing the respectively aforementioned parameters. The data on ACT scores and socioeconomic factors is sourced from EdGap.org and covers nearly 8,000 US schools from the 2016-2017 school year. EdGap is an organization dedicated to mapping the disparity in education outcomes within the US (Memphis Teacher Residency, n.d.) The other two datasets on school type and Title I status feature descriptive statistics on over 100,000 schools in the United States, also collected over the 2016-2017 school year (National Center for Education Statistics, n.d.). They are sourced from the National Center for Education Statistics (NCES). Each dataset was imported into a Jupyter Notebook file, filtered for the factors of interest, and subsequently merged into a clean dataset before further analysis. While the approach utilized in this project is largely descriptive, the study of which factors most greatly influence ACT scores could provide

meaningful insight into where students may require socioeconomic or educational assistance to produce better academic outcomes.

Theoretical Background

The project required the use of Python programming techniques and various Python libraries for data frame manipulation, plotting, and modeling. The modeling process exclusively utilizes multiple linear regression to model the linear relationship between multiple numerical variables. The strength of correlation coefficients derived from these methods will be used to identify the strongest predictors of average ACT scores.

Procedure

The three relevant datasets were imported into a Jupyter notebook and converted into their own respective Pandas DataFrame. The variables of interest, both categorical (charter school, Title I school) and numerical, were filtered to be the exclusive columns within each dataset. All three datasets possessed identical school ID codes that would allow for them to be joined. The NCES datasets were joined first using a left join, followed by another left join to merge with the EdGap dataset.

Following the merging of all three datasets, all rows containing impossible values (negatives, values falling below the correct range) within the shared dataset were deleted for greater clarity. Further care was taken to simplify the Title I column to only include two categorical values, either that the school was not a Title I school or that it received/was eligible for Title I funding. Additionally, prior to exploratory data analysis, missing values within the columns for numerical predictors were imputed. This was accomplished using the 'IterativeImputer' tool from the scikitlearn library, which imputes data estimated from the

relationship between multiple variables. The cleaned and merged dataset was then exported and re-imported as the primary dataframe of interest for exploratory data analysis.

To first interpret the strength of the relationship between average ACT scores and the numerical predictor variables (e.g. unemployment rate, median income, etc.), a correlation matrix was generated using tools from the matplotlib library (Figure 1). The correlation matrix lists the correlation coefficient of each numerical variable relative to one another on a scale of -1 to 1. The relative magnitudes of each correlation coefficient relative to average ACT score was noted prior to modeling.

To further investigate this relationship, statistical modeling was required both to assess the statistical strength of these correlative relationships and determine the impact of the categorical variables. Using statistical modeling tools from the statsmodels library, six separate multiple linear regression models were generated. The first two were designed to model the regression of all numerical predictor variables relative to average ACT score and identify only those with statistically significant coefficients. This process was repeated with the categorical variables of whether the school was a charter school or a Title I school. Each was modeled separately to be paired alongside the numerical variables and assess their correlation coefficients and p-values. The two models of each were also compared based on their mean absolute error, which would provide insight into which model was better representative of the correlative relationship.

Scaling was then applied to each of the numerical predictor variables to alleviate the differing scales for each of their correlation coefficients. This ensured that the distribution of each variable possessed identical means and standard deviations, and allowed for the correlation coefficients of each to be compared according to their magnitudes.

Results

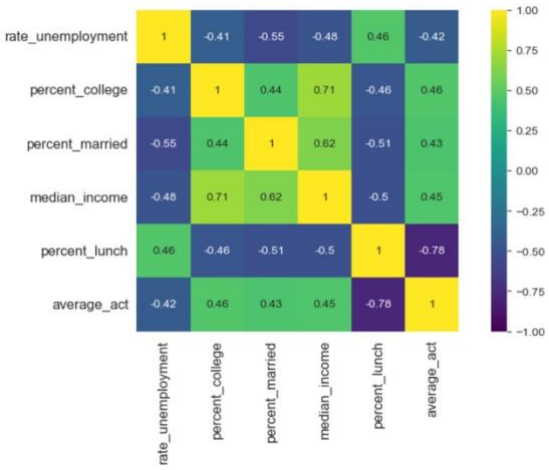

Figure 1. Correlation matrix comparing average_act to all other numerical predictor variables from the dataframe

From the correlation matrix (Figure 1), each numerical predictor possessed a non-zero correlation coefficient relative to average ACT score, with the percentage of students on free or reduced-price lunch possessing the coefficient of the highest magnitude. This result was also exemplified in the model summary of the scaled data (Figure 2). Once each of the three statistically significant numerical variables were scaled, their correlation coefficients conveyed both whether the relationship with average ACT scores was negative or positive, and to what magnitude. The coefficient for 'percent_lunch' was of a much higher magnitude than the other two statistically significant predictors. With respect to the two categorical variables, their model summaries conveyed they had a statistically significant but minuscule impact on the data. However, it should be noted that the largest correlation coefficients for either variable were of the highest magnitude with 'percent_lunch' as well.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.3409	0.019	1082.093	0.000	20.304	20.378
rate_unemployment_normalized	−0.1371	0.022	−6.265	0.000	−0.180	−0.094
percent_college_normalized	0.2667	0.022	12.168	0.000	0.224	0.310
percent_lunch_normalized	−1.7753	0.022	−79.057	0.000	−1.819	−1.731

Figure 2. Excerpt of model summary of normalized multiple linear regression model (includes correlation coefficients)

Discussion

The correlation matrix provided the earliest answer into the strength of the relationship between percentage of students on free or reduced lunch and average ACT scores. Being of the highest magnitude compared to all other variables, the negative coefficient also implied a negative correlation relationship between the two variables, with ACT scores decreasing relative

to increased percentage of students on a lunch program. This was further evidenced following the process of scaling, where although the rate of unemployment and the percent of those with a college degree were significant predictors, the percentage of students on free or reduced lunch had a much higher correlation coefficient.

Furthermore, it was initially hypothesized that eligibility for free or reduced lunch would convey a strong association with the variables of median income and Title I status. This, however, proved not be evidenced by either the exploratory data analysis or the statistical modeling. This could be due to the median income data being sourced from census data, which could skew toward individuals and families without children at the school in question. The Title I status column was also simplified to be a two-pronged variable (either Title I or not), which treated schools eligible for Title I funding as though they were Title I. If reduced lunch increases with schools that are formally Title I schools, this change could dampen a much stronger association that may exist between the two. Additional limitations in the association could also be due to the limited number of schools provided in the EdGap data.

Conclusion

The Education Project data analysis proved to establish the strength of association between average ACT scores and socioeconomic factors. The most predictive variable for a change in ACT score was identified as being the percentage of students on free or reduced lunch. Given that both these parameters examined the same population of students, as opposed to being impacted by census data, their association can be assumed to be highly insightful. The analysis can provide the groundwork for prescriptive proposals on how to improve ACT scores for students on free or reduced lunch or investigate alternatives to standardized testing for all students, not simply those of economically disadvantaged backgrounds.

References

Memphis Teacher Residency. (n.d.). *EdGap: Visualizing the education gap*. https://www.edgap.org/#5/37.718/-95.99

National Center for Education Statistics. (n.d.) *Common Core of Data: Public Elementary/Secondary School Universe Survey Data*. U.S. Department of Education. https://nces.ed.gov/ccd/pubschuniv.asp