Alex Song

Education Project

DATA 5100

Abstract

This project examines relationships between average high-school ACT scores and a set of school-level funding and socioeconomic variables using the analysis coded in https://github.com/alexsong-lab/DATA5100/blob/main/education/code/Education.ipynb. I cleaned and merged EdGap, School Information, which were provide in-class, and 2017 Public Elementary-Secondary Education Finance Dataset from census.gov (note: Education Finance Dataset is indicated as school_funding in the code). The imputation formula was provided by the instructor as a part of pre-class and in-class practices, this formula allows me to impute any missing numeric values. Throughout the analysis, the project tries to explore distributions and correlations and estimated univariate and regression models (socioeconomic, funding). Key findings are socioeconomic variables (median household income, percent adults with college degrees, percent free/reduced lunch, unemployment) explain the largest share of variation in school-level ACT averages ($R^2 \sim 0.63$), while individual funding variables (per-pupil total expenditure, per-instructor salary) show statistically significant but much smaller when modeled alone ($R^2 \sim 0.004$ and 0.034) and are highly collinear with each other ($r \sim 0.89$).

Introduction

Education performance such as ACT scores are affected by many factors, including school resources and socioeconomic indicators. This project tries to understand whether the education performance e.g., average ACT score outcomes are primarily driven by socioeconomic conditions, by school resources and funding, or by the interaction of those factors. This report analyzes a merged dataset that combines school-level ACT score, socioeconomic indicators, and school funding indicators. The goal is to explain and identify relations in which predictors carry the most reliable explanation for average ACT scores at the school level.

Theoretical background

Linear regression is a simple, widely used tool to measure how an outcome changes with predictors while holding other measured variables constant. Regression estimates describe associations, not causation. Also, if several variables move very similarly, e.g., PerPupil_Total_Expenditure and PerInstructor_Salary shown in the project, it becomes hard to

trust how important each one is on its own. Socioeconomic factors like local income and adult education often affect student preparation for ACT score, and school financial factors tend to change together across places. Please note that EdGap data and socioeconomic variables are provided at the school level, while the school financial variables are aggregated at the district level. Due to the difference in the levels of measurement, some of the visualizations may introduce distorted interpretation and bias. Because district totals may not reflect resource differences across individual schools (i.e., every single school within the same district), any analysis or results provided in this report that combine these sources should be interpreted with extra caution.

Methodology

Data preparation and selection follow these steps:

- Started by combining three dataset so each row represented one high school with its ACT average, socioeconomic factor, and district funding information where available. The files were EdGap (school outcomes and socioeconomic indicators), the School Information file provided in class, and the 2017 Public Elementary-Secondary Education Finance dataset (the notebook calls this school_funding). To keep the outcome comparable, I kept only schools labeled as High and converted clearly invalid values (for example, implausible ACT scores) to missing before removing those rows.
- Renamed columns to more easy-to-understand and readable labels (for example PPCSTOT → PerPupil_Total_Expenditure) and matched school and district IDs to merge/join the datasests. Note: because the finance data are reported at the district level while the outcomes and many socioeconomic factors are at the school level, I noticed that this mismatch could affect interpretation, and misinformed variable values sometimes.
- Missing data were minimal for EdGap socioeconomic indicators (under 1%) but some school funding indicators had many zero values (about 5–6%). To avoid losing many schools, I used an IterativeImputer to fill in numeric gaps, which preserves sample size and maintains relationships among variables to without negative affecting the analysis.

The code and statistics used:

- Exploratory checks such as summary statistics, correlations, and scatterplots with linear and quadratic fits. I also paid attention to correlations among funding indicators; for example, PerPupil_Total_Expenditure and PerInstructor_Salary are very highly correlated ($r \sim 0.89$), therefore, their scatterplot looks extremely similar, almost identical. Hence, they carry overlapping information.
- For the main analysis I used regression models. I began by running simple one-variable regressions to see how each factor related to ACT scores. For median income, I added a squared term to check if the relationship curved rather than being a straight line. After

that, I built multiple variable models that combined several socioeconomic predictors together.
- For multiple linear regression models following were used to standardize the predictors: r2, mean absolute error, and root mean squared error.

To summarize, I checked that the model satisfied its basic assumptions, interpreted the coefficients as the average change in the ACCT scores, and placed more emphasis on the overall look of the model rather than on individual coefficients when predictors overlapped. The reason for this was to examine common issues that often arise in regression models. For example, I looked for problems such the presence of outliers, or predictors that were too strongly correlated with each other.

Computational results

The dataset included about 7,200 high schools (around 7,900 before removing schools with missing values). When I looked at the link between average ACT scores and median household income, income alone explained about 21% (.211) of the differences across schools (OLS Regression Result). The prediction error was about 2.2 ACT points on average. Adding a squared term for income improved the accuracy just a little, suggesting the relationship is not perfectly straight but slightly curved.

School funding by itself explained very little. For example, per-pupil spending explained less than 1% of the variation, and average instructor salary explained only about 3%. These funding indicators are also very similar to each other. It seems districts that spend more per pupil also tend to pay higher salaries, or maybe it's just missing individual school data, so it's hard to separate their individual effects. A multi-lineal and reduced model explained about 63% (0.628) of the variation in ACT scores. This was almost the same performance as larger models with many more variables.

Discussion

What is apparent in this analysis is that socioeconomic factors such as household income, adult education levels, and student economic need are the strongest predictors of average ACT scores. A few well-chosen socioeconomic indicators explain most of the differences between schools, so a small, simple model works almost as well as a larger one.

School funding indicators also show some relationship with ACT scores when looked at on their own, but once socioeconomic conditions are considered, their added value is very limited. Because funding indicators tend to move together, it's difficult to say which type of financial factor has an independent effect. In other words, differences in funding often reflect differences

in community wealth, so the link between money and scores may be more about context than direct impact.

Conclusion

Putting the pieces together, socioeconomic indicators are the most reliable and powerful predictors of school-level average ACT scores in these merged data. Financial measures such as per-pupil spending or per-instructor salary also show some association with scores, but they are highly correlated with one another. Moreover, these measures are directly tied to community wealth, such as median household income, and to other socioeconomic factors. For instance, schools in higher-income areas are more likely to have higher per-pupil expenditures and instructor salaries. Thus, when these variables are considered together, financial indicators may provide some additional explanation.