**Education Project Report**

**Abstract:** Predicting School Performance from Socioeconomic Factors Using Multiple Linear Regression

**Author:** Badamgarav Battushig

**Course:** DATA 5100 – Foundations of Data Science

**Instructor:** Professor Brian Fischer

**Institution:** Seattle University

**Date:** October 22, 2025

In this Education Project, we will consider how socio-economic factors affect students’ ACT scores. There are multiple factors that could be the main drivers of students’ ACT performance. For this project, I use multiple linear regression because ACT score is a number that can be any value between 1 and 36, so we can use linear regression to study it. Linear regression helps us see if there is a straight-line relationship between the ACT score and many things at the same time. In this project, we divide the socioeconomic part into:

1. Economic factor: Household income
2. Poverty factor: Percent of students on free or reduced lunch
3. Employment stability: Unemployment rate

Before we begin, we tested all three parts in the demo using two datasets: *EdGap* and *School Information*. This time, we used three datasets by adding the *Public School Characteristics* (renamed as “school”) dataset. This dataset includes the school ID, school level, lunch-related information, teacher-student ratio, and demographic data.

I am interested in exploring how school staff influence student achievement. Teacher availability reflects school resources, so it is important to see how this affects student performance.

When merging datasets using the common school ID (id) field, each dataset contained school-level information. The merge was performed using a left join to keep all records from the main dataset. Since our scope focuses on ACT scores, we only need high school level data. The new dataset also included school level information, so we filtered it first, dropped unnecessary columns, and then merged the files.

For data cleaning, several variables contained invalid or inconsistent values that could distort the analysis.

* percent\_lunch included negative percentages and values exceeding 100%, while average\_act contained unrealistic scores below 1. These were replaced with missing values and later imputed using column means.
* The percent\_lunch column was rescaled from decimals to percentages for better interpretation, and the dataset was filtered to include only high schools.
* The funding variable (“No/Yes”) was recoded into binary form (No = 0, Yes = 1), and missing values were filled using the most frequent category.

Economic Factor

In this section, we use census tract data for median household income, percentage of adults with a college degree, percentage of children in married-couple families, and average ACT scores.

The heatmap shows positive trends between household income and test scores. The correlation matrix uses coefficients between -1 and +1:

* Bright yellow (close to +1) means a strong positive relationship: when one variable increases, the other increases.
* Dark purple (close to -1) means a strong negative relationship: when one increases, the other decreases.
* A colorful squares with white text

  AI-generated content may be incorrect.Greenish colors near 0 indicate little or no relationship.

A chart of a graph

AI-generated content may be incorrect.In the heatmap, we see consistent positive trends between test results and economic well-being. The pairplot also shows upward trend lines, confirming a positive relationship. We can say that community wealth improves students’ ACT performance.

Poverty Factor

This section includes the percentage of students eligible for free or reduced-price lunch, Title I program participation (federal funding for low-income schools), number of students receiving free lunch, and ACT scores. We already know that lunch variable is a strong predictor of ACT scores. Overall, economic disadvantage and limited school funding are the strongest predictors of lower ACT performance.

Employment Stability

This section includes unemployment rate, full-time equivalent teachers, and student–teacher ratio. The correlation matrix shows that community stability influences student performance. This suggests that external economic factors play a larger role on students’ test scores.

Multiple Linear Regression Model

The multiple linear regression model examines how socioeconomic and school factors collectively predict average ACT scores. The independent variables include income, unemployment rate, education level, lunch participation, and funding.

We focus mainly on the p < |t| value, which shows statistical significance. When p < 0.05, the variable is considered significant, meaning that its effect on ACT scores is unlikely due to chance. In this model, almost all predictors such as income, percent of students receiving free/reduced lunch, unemployment rate, and education level—are significant and strong predictors. However, variables like *percent\_married* with p > 0.05 are not considered strong predictors.

A screenshot of a computer screen

AI-generated content may be incorrect.The R² and Adjusted R² values show model strength. The model’s R² = 0.63 means that 63% of the variation in ACT scores is explained by these predictors. High R²/Adjusted R² values with low p-values indicate that the regression model fits the data well and includes meaningful predictors of ACT scores.

Reduced Model

The reduced multiple linear regression model includes only the strongest and most influential predictors of student performance. The R² = 0.623, meaning 62.3% of the variation in ACT scores is explained by these predictors. All predictors have p-values < 0.001, confirming that each significantly affects ACT scores.

A screenshot of a computer screen

AI-generated content may be incorrect.This model clearly shows that poverty, employment, and economic factors are the strongest socioeconomic predictors of ACT outcomes, even when other variables are excluded.

Normalized Model

To make predictor effects directly comparable, the same reduced model was re-estimated after normalizing the continuous predictors. Normalization scales each variable to have a mean of 0 and standard deviation of 1, allowing us to interpret coefficients as standardized effects.

A screenshot of a computer

AI-generated content may be incorrect.The normalized model produced identical fit metrics (R² = 0.623, MAE = 1.1411). This shows that scaling does not affect model accuracy but makes it easier to compare predictor importance.  
Both reduced and normalized models confirm that economic disadvantage, funding, and community employment are the most critical factors influencing student ACT performance.

Therefore, this project shows that socioeconomic factors (including school staffs) have a big effect on how well students do on the ACT test. When families have more money and parents have good jobs, students tend to get higher scores. On the other hand, when families do not have enough money get lower scores. The school also needs enough money and teachers to help students learn well. From our study, we found that 63% of student ACT scores can be explained by these socioeconomic factors. The results of this project help us to understand what we need to do to help all students have a fair chance to succeed.

**Reference List**

**EdGap.org dataset:**

Fischer, B. (2023). *EdGap.org Educational Data Portal* [Data set]. GitHub repository:  
https://github.com/brian-fischer/DATA-5100

**NCES School Information dataset:**

National Center for Education Statistics. (2023). *Common Core of Data (CCD) – Public School Universe Survey Data* [Data set]. U.S. Department of Education.  
https://nces.ed.gov/ccd/schoolsearch/

**Public School Characteristics:**

National Center for Education Statistics. (2021). *Public School Characteristics, 2020–2021* [Data set]. U.S. Department of Education.https://catalog.data.gov/dataset/public-school-characteristics-2020-21-e279c/resource/267cd902-7c6c-4121-93f7-e089e205a95c