

Final Project: College Data Analysis

STAT 135 (Intro to Stats via Modeling) Project Report

Group A: Nikol Zyzen and May Mishima

2025-05-06

Table of contents

Abstract	1
Introduction and Goals	1
Description of Dataset and Variables	2
One Graphical Display	3
Multiple Linear Regression Model and Interpretation	3
Diagnostics for the Model	6
Conclusion and Takehome Message	8

Abstract

This project explores how college characteristics relate to success after graduation, measured by median earnings ten years after enrolling. We focused strictly on New England institutions, analyzing how their admission rates and graduation rates correlate with future earnings. We found that schools with higher graduation rates and lower admission rates tend to have graduates with higher earnings, especially among private nonprofit institutions. However, these findings are limited by assumptions of the linear regression model and the regional scope of the data, so readers should interpret results with caution given these constraints. Still, the study offers valuable insight into identifying measurable indicators of college effectiveness and long-term student success.

Introduction and Goals

Choosing which college to attend is one of the most important decisions many high school students face. Although there's a wealth of information available online, it's not always clear which statistics are most relevant. This project explores the factors that may contribute to student success over time, measured by average income ten years after enrollment. If you're

interested in understanding the earnings potential associated with different colleges, this study can help highlight which data points are worth paying attention to. The findings may also prove useful for developing or validating college ranking systems and for colleges seeking to identify which factors to prioritize in order to improve their students' long-term success. Data comes from the College Scorecard, which aggregates information at the institutional level.

Description of Dataset and Variables

Our dataset is 443 4-year New England institutions from the US Department of Education's College Scorecard. Each observation represents an individual school and contains variables related to admissions, student outcomes, and institutional characteristics.

Table 1: Summary of key variables from the College Scorecard dataset used in the study

Variable	Description	Note
Mn_earn_wne_p10	Median earnings of students 10 years after entering.	The median earnings are right skewed, with a median of approximately \$41000, and the range of the middle 50% of observations is about \$22000. There are a couple of high-end outliers exceeding \$100000 in earnings.
Adm_rate	Admission rate or proportion of applicants who were admitted.	The distribution of admission rate is left-skewed, with a median of around 71% and the range of the middle 50% of observations around 26%. A few institutions have very low admission rates, so there are outliers for the highly selective schools on the lower end.
C150_4	4-year graduation rate, measured as completion within 150% of the expected time.	The distribution of the 4-year graduation rate is relatively symmetrical, with a mean of around 58% and a standard deviation of 21%, meaning that, on average, graduation rates for schools in this dataset fall 21% above or below the mean. There are some college outliers that have graduation rates below 25%.

Variable	Description	Note
Control	Type of institution– private nonprofit, private for-profit, or public.	In this dataset, public and private nonprofit institutions make up the majority, and private nonprofits make up the smaller portion.

Table 1 displays the codebook for our data. Our dataset consists of $n = 443$ observations.

One Graphical Display

Figure 1 shows the relationship between 4-year college completion rates and median income 10 years after enrollment. Higher completion rates generally correlate with higher incomes, especially among private nonprofit institutions. Colleges with average incomes above \$100,000 are labeled—most are elite private nonprofits, with one exception: Maine Maritime Academy, a public college with a comparatively lower graduation rate.

Figure 2 displays the relationships among median income, admission rate, and 4-year enrollment rate, grouped by institution type. Higher incomes are associated with lower admission rates and higher graduation rates, particularly among private nonprofit colleges. Public and for-profit institutions generally show weaker correlations and lower income outcomes.

Multiple Linear Regression Model and Interpretation

Table 2: Multiple Linear Regression

term	estimate	std.error	statistic	p.value
(Intercept)	50193	10872	4.617	0.000
adm_rate	-30818	7594	-4.058	0.000
c150_4	53065	9301	5.705	0.000
controlPrivate nonprofit	-6895	7036	-0.980	0.329
controlPublic	-4634	7176	-0.646	0.519

Table 2 displays the regression results. The multiple regression model estimated is:

$$\text{Income}_i = \beta_0 + \beta_1 \cdot \text{AdmissionRate}_i + \beta_2 \cdot \text{GradRate}_i + \beta_3 \cdot D_{i,\text{Private_nonprofit}} + \beta_4 \cdot D_{i,\text{public}} + \varepsilon_i$$

The model has an R-squared of 0.522, indicating that approximately 52.2% of the variation in income 10 years after enrollment is explained by the included predictors. The standard error

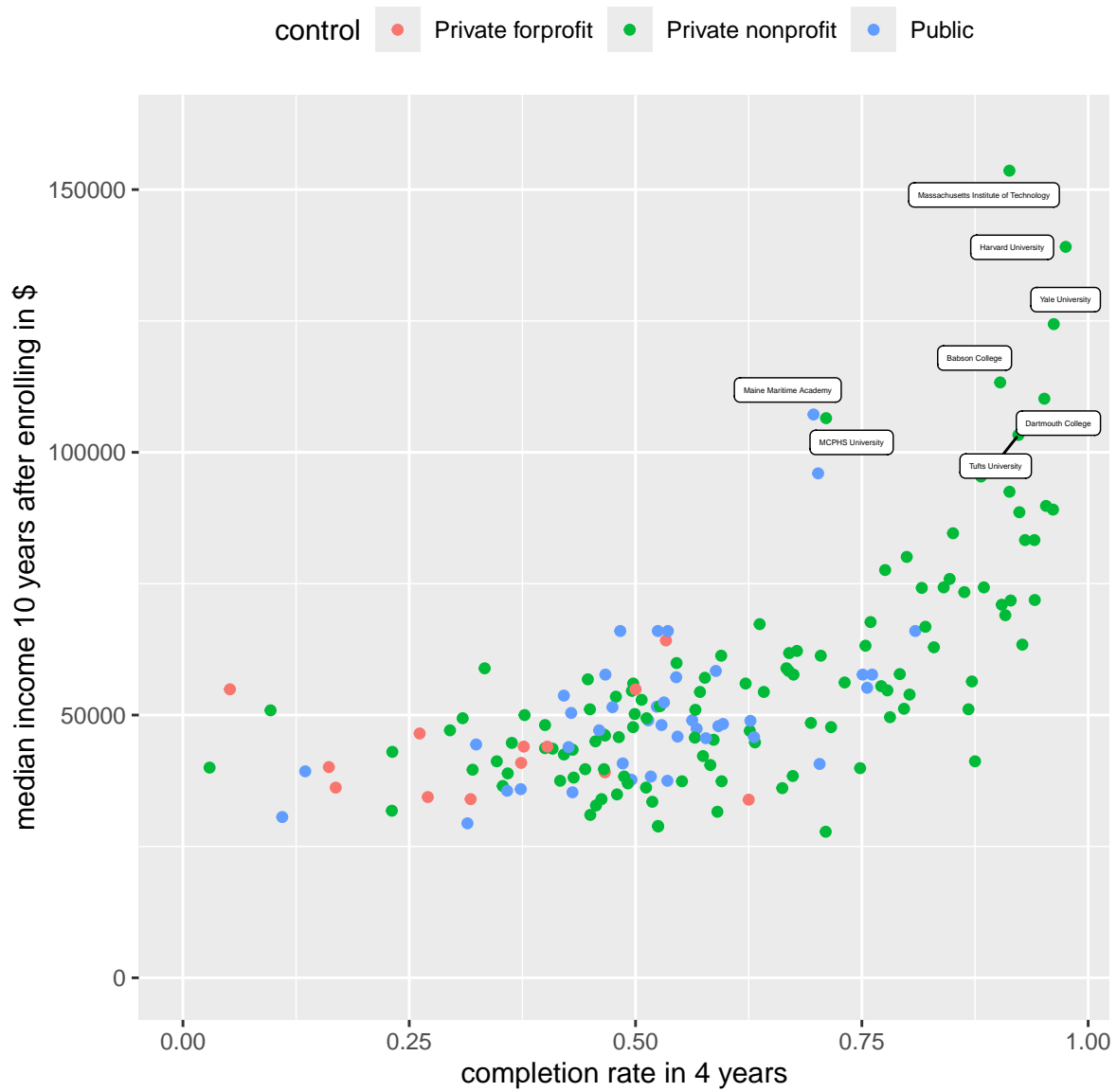


Figure 1: 4-year college completion rates and median income 10 years after enrolling

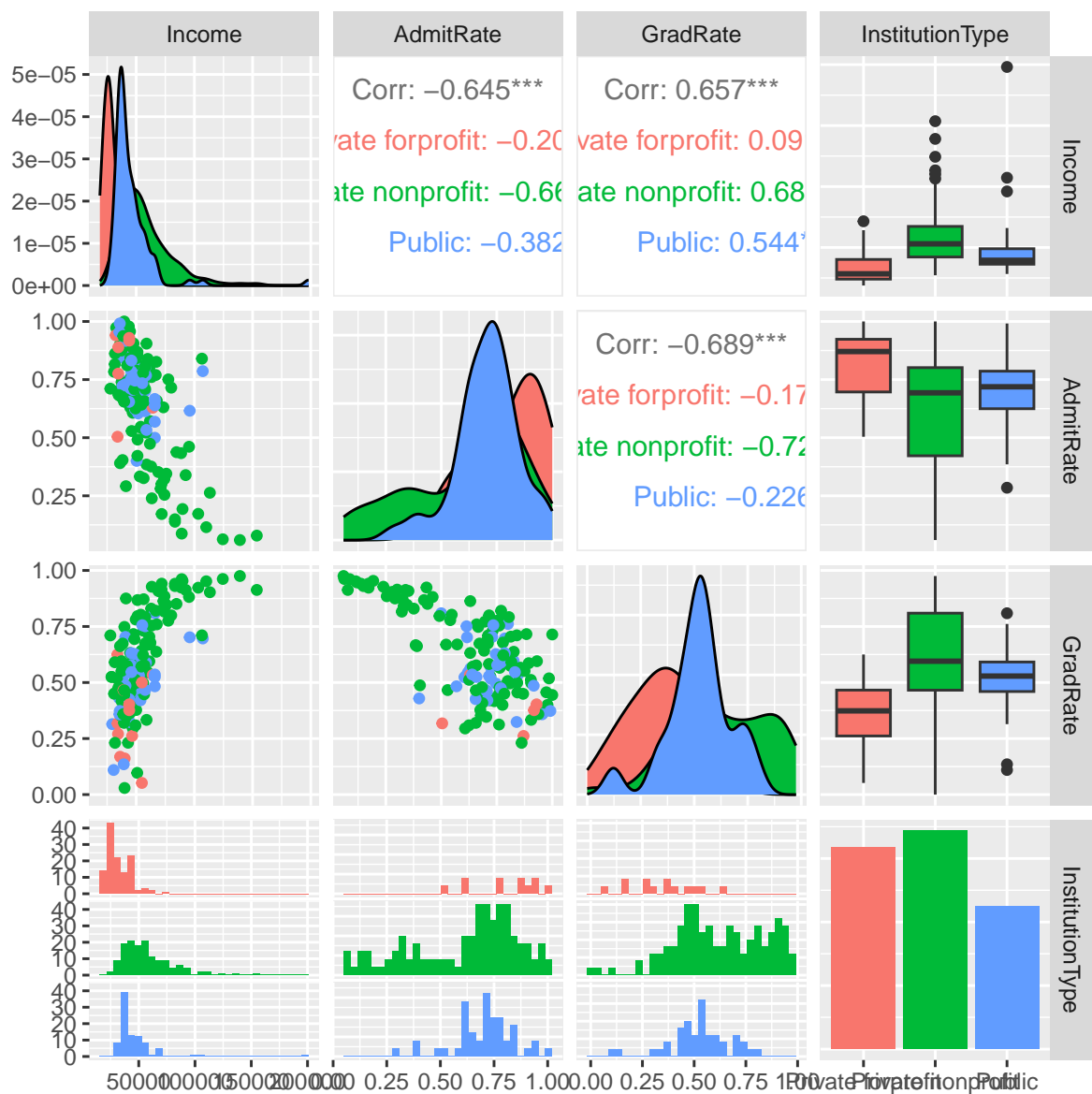


Figure 2: pairs plot

of 14,800 suggests that, on average, actual incomes deviate from the model's predictions by about \$14,800.

The model is statistically significant overall, with an F-statistic of 39.1 on 4 and 143 degrees of freedom and a p-value < 0.001 . This large F-value and very small p-value indicate that at least one predictor contributes meaningfully to explaining income variation.

The intercept is \$50,194, representing the predicted income for someone attending a for-profit institution with both a 0% admission rate and a 0% graduation rate, which is obviously an unrealistic scenario.

The coefficient for admission rate is -30,818, meaning that a 1-unit increase in admission rate is associated with a \$30,818 decrease in median income, on average. This relationship is statistically significant ($|t| = 4.06$, $p < 0.0001$).

The coefficient for 4-year graduation rate is 53,066, indicating that a 1-unit increase in graduation rate is associated with a \$53,066 increase in income, and this effect is also statistically significant ($t = 5.71$, $p < 0.0001$).

Institution-type effects show that compared to for-profit colleges, attending a private nonprofit institution is associated with a \$6,895 decrease, and attending a public institution with a \$4,634 decrease in income. However, neither coefficient is statistically significant, meaning these differences could be due to random variation.

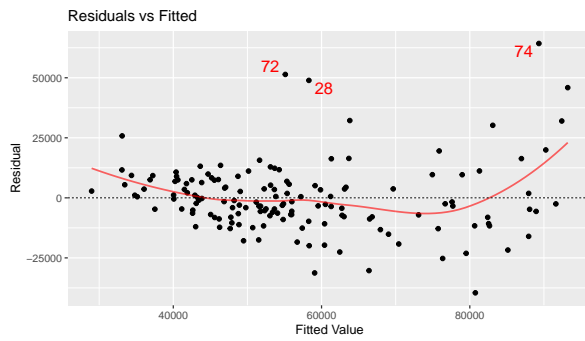
In summary, the model suggests that higher admission rates are associated with lower post-graduation income, while higher graduation rates are associated with higher income, and both effects are statistically significant.

Diagnostics for the Model

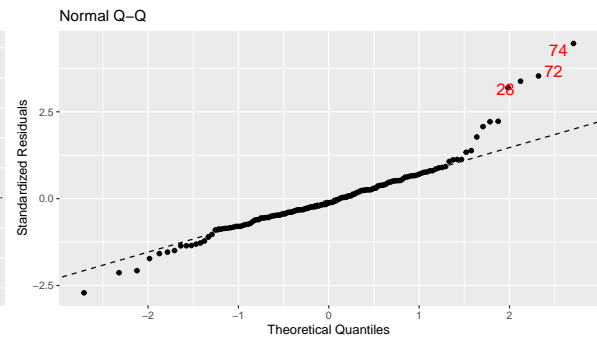
Figure 3 displays the regression diagnostics for the model.

When running a multiple regression model, we want to check the assumptions of linearity, independence, constant variance, and normality. We assume the independence condition is met, given the nature of the data representing individual institutions. The Q-Q plot shows noticeable deviations from the diagonal, especially in the upper tail, so the normality assumption is not satisfied. To assess equal variance and linearity, we examined the Residuals vs Fitted and Scale-Location plots. The increasing spread of residuals suggests a violation of the constant variance assumption. Additionally, the curved pattern in the Residuals vs Fitted plot indicates potential non-linearity, suggesting that at least one predictor may not have a linear relationship with the outcome.

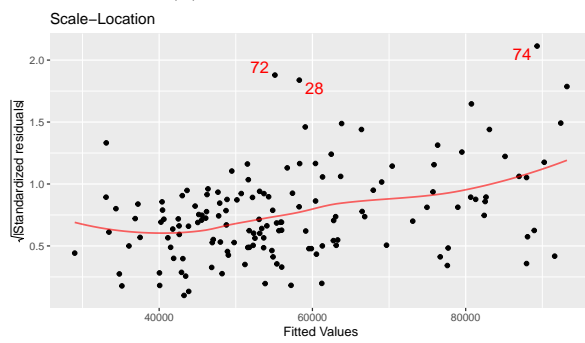
A few notable outliers include Maine Maritime Academy, the University of Maine at Machias, and Manchester Community College. These points produce unusually large residuals: Maine Maritime Academy reports a very high income for a public institution (\$107,200), while the



(a) Residuals vs Fitted



(b) Q-Q plot



(c) Scale-Location Plot

Figure 3: Model diagnostics

other two have exceptionally low incomes (\$29,400 and \$36,600), making them stand out from the overall trend.

Conclusion and Takehome Message

We found a statistically significant relationship between admission rate and average earnings after 10 years, as well as between graduation rate and average earnings after 10 years. Higher admission rates were associated with lower earnings, while higher graduation rates were linked to higher earnings—particularly among private nonprofit institutions. However, it's important to note that our analysis focused only on New England schools, which may limit the generalizability of our findings. Additionally, violations of the linearity, normality, and equal variance assumptions make us cautious in interpreting the results.