Predicting post-graduate earnings with a focus on diversity

Abstract: Earning a college education in the U.S. has long been touted as an opportunity for significant socioeconomic mobility and change, but is this true across all demographic groups? The purpose of this paper is to build a model to predict mean earnings post-graduation based on a variety of factors, with a particular focus on demographic factors. Using data from the governmental database College Scorecard, my multiple linear regression model found that among the nine significant predictor variables, there is a significant negative linear association between proportion of low-income students and mean earnings, and there is a significant positive linear association between proportion of male faculty and mean earnings. It also found that there was a significant relationship between mean earnings and the interaction between highest degree awarded and completion rate of white students.

Background and Introduction

Earning a college education in the U.S. has long been touted as “a panacea for social and economic ills,” an opportunity for significant socioeconomic mobility and change [1]. By 2016, over 95% of the jobs created post-2008 Great Recession were filled by workers with at least some college education [2]. However, there is a documented history of inequity in higher education: A 2019 Education Trust study found significant racial differences in degree attainment, with 47% of white adults holding as associate’s degree or higher compared to 30.8% of Black adults and 22.6% of Latinx adults—meaning that those additional jobs are disproportionately going to white adults rather than adults from minority groups [3]. Even when we do consider just college-educated individuals, there is still significant disparity within that population, as a 2016 Pew Research Center Study found that white men outearn most other demographic groups, even when controlling for education [4].

In the past few years, some companies have begun to shift away from a four-year college degree requirement. In 2017, 51% of jobs required a degree, and by 2021, that figure had dropped to 44% [5]. As the job market continues to evolve, I’m interested in seeing whether post-graduate earnings can be predicted in a more scientifically rigorous manner, and, if diversity in education impacts those earnings, how so and to what degree. Does having a college degree really matter for one’s future salary, and if so, what factors determine how large or small that salary is?

Methods

The dataset was taken from the U.S. Department of Education College Scorecard’s most recent institutional-level dataset released on Oct. 10, 2023 [6]. Because this is meant to be a survey of all institutions of higher education in the U.S. that is conducted by the federal government, it seems reasonable to assume that this dataset is representative of the population. The data are also independent, as one college’s measurements are not influenced by another’s. The original dataset contained the measurements of 3232 variables from 6543 institutions of higher education across the U.S. Most of these variables were variants (e.g. mean earnings of students working and not enrolled 6, 7, 8, 9, and 10 years after entry) or demographic breakdowns (e.g. 3-year repayment rate for low-, middle-, and high-income students) of the same factors. I selected 33 initial predictor variables (14 of which are related to demographics) as seen in Appendix A. After performing univariate and bivariate analyses (some of which can be seen in Appendix B) and collapsing categories with small sample sizes, I settled on 11 predictor variables to include in my model: region (South, West, Midwest, East, West, Outlying Territories), locale (City, Rural, Suburb, Town), highest degree awarded (None/Certificate, Associate’s, Bachelor’s, Graduate), type of institution (Public, Private Nonprofit, Private For-profit), proportion of students who are low-income (aided students whose family income is between $0 and $30,000), average faculty salary, average cost of attendance, institutional expenditure per full-time student, average SAT-equivalent score of students admitted, proportion of full-time faculty who are men, and completion rate of white first-time, full-time students (completion meaning 150% of expected time to completion). After removing missing values, our final sample size was reduced to 3090 observations.

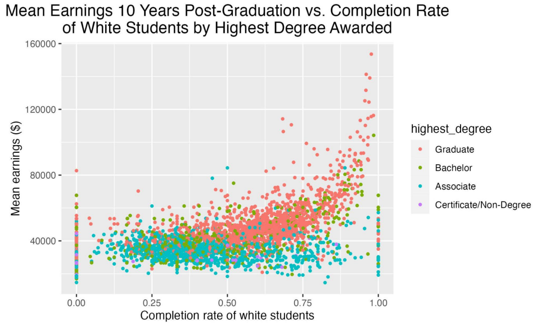


Figure 1. Scatterplot of mean earnings 10 years post-graduation in dollars vs. completion rate of white students

I fit an initial multiple linear regression model (1) to predict mean earnings of students working and not enrolled 10 years after entry from the aforementioned 11 predictor variables and an interaction term between completion rate of white students and highest degree awarded due to the potential interaction indicated by the split tail in Figure 1. Diagnostics for model (1) as seen in Appendix C demonstrate that there are concerns about the Normality and homoscedasticity assumptions for a multiple linear regression model, so I performed a Boxcox power transformation of lambda = -0.2 on mean earnings and removed all variables that were insignificant (α = 0.05) under both a t-test and a partial F-test. My final model (2) predicted the transformed mean earnings from region, locale, highest degree awarded, type of institution, proportion of low-income students, average faculty salary, proportion of male faculty, completion rate of white students, and the interaction between highest degree awarded and completion rate of white students. Diagnostics for model (2) also seen in Appendix C show that there are no more concerns about heteroscedasticity, and while there are still a few trailing points in the Normal Q-Q Plot (Figure 12), the points look much more linear than before. These points of concern may potentially be outliers—however, because the residual plot (Figure 11) and added-variable plots to test for linearity (Figures 13-16) do not provide much cause for concern about these outliers, I decided not to remove the outliers from the dataset and kept model (2) as my final model.

Results

My final model (2) explains 75.77% of the variability in expected mean earnings raised to the -0.02 power (adjusted $R^2 = 0.7577$). With an F-statistic of 505.4 and a p-value of $2.2 \cdot 10^{-16}$, we have enough evidence to reject the null hypothesis that there is no association between any of my predictors and the transformed mean earnings and conclude that model (2) is a better fit for the data than an intercept-only model.

Because I performed a negative transformation, all the coefficients seen in Appendix D are actually flipped in sign—for example, every one dollar increase in average faculty salary is correlated with a $1.24 \cdot 10^{-6}$ decrease in expected mean earnings raised to the -0.02 power when holding all other variables in the model constant, which in fact means that mean earnings is expected to increase, though the coefficient is negative. As we can see from this example, interpretability of this model is complicated due to this negative power transformation—however, we are still able to extract some more general inferences and patterns. This model tells us that as average faculty salary increases, so too does expected mean earnings, and when compared to a public institution, both private nonprofit and for-profit institutions expect to see an increase in mean earnings when holding all other variables constant. Of particular note due to our research interest in demographic variables is that mean earnings is expected to increase when the proportion of low-income students decreases and when the proportion of male faculty increases, if all other variables are held constant. And when we consider the interaction between completion rate of white students and highest degree awarded, we know that when compared to institutions that offer graduate degrees, institutions that only offer associate's degrees (which are the same in all other respects) have lower expected mean earnings (positive coefficients for completion rate of white students, the highest degree category of associate's, and the interaction between completion rate of white students and highest degree of associate's), but we do not know for sure for the other highest degree categories due to differing coefficient signs between all three relevant variables.

It is important to note that completion rate of white students is not significant under either a t-test or a partial F-test, as seen in Appendix D. However, I made the decision to keep it as a variable in my final model because the interaction between completion rate of white students and highest degree awarded is significant, and considering my exploratory data analysis as seen in Figure 1, this interaction seems important to a model predicting mean earnings. When including an interaction in a model, it is important to keep the main effects in the model as well, even if one of them is not itself significant.

Discussion

My final statistical model shows that post-graduate earnings can be predicted, not only on certain attributes of the institution of higher education itself but also on a couple of demographic variables. Unsurprisingly, based on a historical trend of inequity in higher education, there is a significant negative linear association between proportion of low-income students and mean earnings of students working and not enrolled 10 years after entry, and there is a significant positive linear association between proportion of male faculty and mean earnings. In this model, while completion rate of white students by itself does not have a significant relationship with mean earnings, it does influence the relationship between highest degree awarded and mean earnings through an interaction term such that it obfuscates what would otherwise be a clear expected relationship of an institution offering higher degrees also having higher mean earnings. Compared to an institution that offers graduate degrees, while we know that an institution offering only associate's degrees will have a lower expected mean earnings, we cannot definitively state the direction of the difference in mean earnings for institutions that only offer bachelor's or certificate's/no degrees.

Based on this findings, it's clear that student and faculty compositions matter when predicting post-graduate earnings, and this relationship is likely tied to historical structures of power in education (and in the world) that favor white men. However, it is important to address the limitations of this model.

1) There were over 3000 missing values that had to be removed from the original dataset. Part of this is due to the fact that there were simply missing measurements in the form of NULL responses—however, there were also many measurements that were not released for privacy reasons, as small sample sizes could reveal identifiable sensitive information to the public. As a result, the concern for this model is less the sample size of the dataset itself (3090 is still quite a large amount of observations) but rather what information is being lost due to these missing values. Perhaps certain institutions are more likely to have measurements be suppressed for privacy reasons (e.g. institutions in rural locations, or institutions in the outlying territories), which means that they are underrepresented in the dataset used to build my model.
2) As seen in Appendix B, Figure 2, there were several high outliers in post-graduate mean earnings that (as mentioned in the Results section) seemed to have an impact on the Normal Q-Q plot in our diagnostics for our final model. It could be helpful to see what predictors would be included in a model fit to the data with these outliers removed—or to consider other factors that could potentially explain these outliers. For example, prestige is a variable that I did not include in the model, in part because it was difficult to find a way of encoding it. At first, I planned on using admission rate—however, there were over 3000 missing admission rate values in the original dataset. Perhaps a future study could incorporate prestige by scraping the U.S. News & World Report's national university rankings. Another variable I did not include was proportion of majors: for example, how many graduates majored in a STEM degree versus a humanities degree. College Scorecard does offer a dataset that specifically disaggregates by field of study, but most of the data in it is suppressed due to privacy reasons.

It may be difficult to address the first limitation completely, as many of the missing values are suppressed for privacy reasons and are thus a matter of ethics. However, the missing values due to data collection could be rectified by a more robust and thorough procedure. I've also mentioned a few ways of addressing the outliers in the dataset: Either remove them or see if they can be explained through variables I didn't include in the model, such as prestige and/or major. Perhaps most important of all is expanding the scope of our current definition of diversity to include other types of diversity such as LGBTQ+ diversity and neurodiversity.

References

1. Kirshner, Jodie Adams. “Op-Ed: Higher education as a path out of poverty is now more myth than reality.” *LA Times*, 29 January 2023, https://www.latimes.com/opinion/story/2023-01-29/higher-education-college-degree-poverty-student-debt-loans.
2. Carnevale, Anothony P., et al. *America’s Divided Economy: College Haves and Have-Nots*. Georgetown University Center on Education and the Workforce, 2016, https://cew.georgetown.edu/wp-content/uploads/Americas-Divided-Recovery-web.pdf.
3. Jones, Tiffany and Katie Berger. *Aiming for Equity: A Guide to Statewide Attainment Goals for Racial Equity Advocates*. The Education Trust, 2019, https://s3-us-east-2.amazonaws.com/edtrustmain/wp-content/uploads/2019/01/08151345/Aiming-For-Equity.pdf.
4. Patten, Eileen. “Racial, gender wage gaps persist in U.S. despite some progress.” Pew Research Center, 1 July 2016, https://www.pewresearch.org/short-reads/2016/07/01/racial-gender-wage-gaps-persist-in-u-s-despite-some-progress/.
5. Lohr, Steve. “A 4-Year Degree Isn’t Quite the Job Requirement It Used to Be.” *The New York Times*, 8 April 2022, https://www.nytimes.com/2022/04/08/business/hiring-without-college-degree.html.
6. “Download the Data.” *College Scorecard*, U.S. Department of Education, 10 October 2023, https://collegescorecard.ed.gov/data/.

Appendices

Appendix A - Variable Selection

Variable	**Description**
mean_earnings_10_years	Mean earnings of students working and not enrolled 10 years after entry
name	Name of the institution
city	City institution is located
state	State the institution is located
region	What region of the US is the institution located in? (South, West, Midwest, East, West, Outlying Territories)
locale	Describes the geographic location of the institution, categorized based on its level of urbanization (City, Rural, Suburb, Town)
highest_degree	Highest degree awarded (Graduate, Bachelors's, Associate, Certificate/Non-degree Granting)
level	Level of institution (4 year, 2 year, less than 2 year)
type	Type of institution (Public, Private Nonprofit, Private For-profit)
women_only	Women-only institution
men_only	Men only institution
religious	Religious affiliation of the institution
PWI	Institution is predominantly white (Yes, No — HBCU, PBI, AANHI, TRIBAL, AANAPII, HSI, and/or NANTI)
median_debt	The median debt for students who have completed their degree
pct_first_gen	Proportion first-generation students
pell_grant	Share of students who received a Pell Grant while in school
avg_fam_inc	Average family income
pct_students_low_income	Proportion of aided students whose family income is between $0-$30,000
pct_loan	Proportion of all undergraduate students receiving a federal student loan
endowbegin	Value of school's endowment at the beginning of the fiscal year
endowend	Value of school's endowment at the end of the fiscal year

expenditure	Instructional expenditures per full-time equivalent student
attendance_cost	Average cost of attendance
avg_fac_sal	Average faculty salary
pct_fac_men	Share of full time faculty that are men
pct_fac_white	Share of full-time faculty that are white
student_fac_ratio	Undergraduate student to instructional faculty ratio
pct_students_men	Total share of enrollment of undergraduate degree-seeking students who are men
grads	Number of graduate students
admission	Admission rate
avg_sat	Average SAT equivalent score of students admitted
completion_rate	Completion rate for first-time, full-time students at four-year institutions and less-than-four-year institutions (150% of expected time to completion)
completion_white	Completion rate for first-time, full-time students at four-year institutions and less-than-four-year institutions (150% of expected time to completion) for white students
retention_rate	First-time, full-time student retention rate at four-year institutions and less-than-four-year institutions
transfer_rate	Transfer rate for first-time, full-time students at four-year institutions and less-than-four-year institutions

These are the 33 initial predictor variables I chose, roughly separated into three groups: 1) Red: general attributes of the institution, 2) Orange: money-related attributes of the institution, and 3) Yellow: student and faculty composition. The bolded variables are the ones that were ultimately included in my final model (2). I removed variables that either had a large amount of missing values (e.g. admission rate) or, based on initial univariate and bivariate analyses, did not seem like they would be as good a predictor as other variables that were attempting to measure the same thing (e.g. choosing proportion of low-income students over proportion of students receiving a federal student loan as a measurement of socioeconomic diversity in the student body).

Appendix B - Exploratory Data Analysis

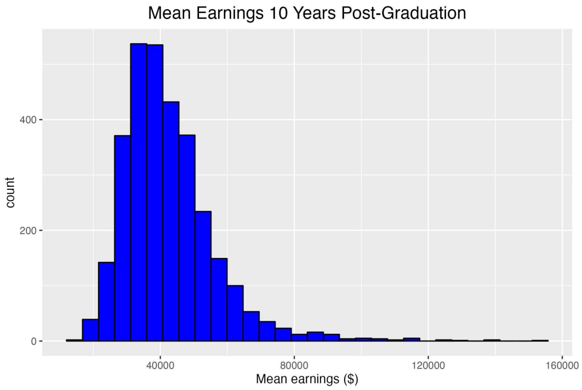
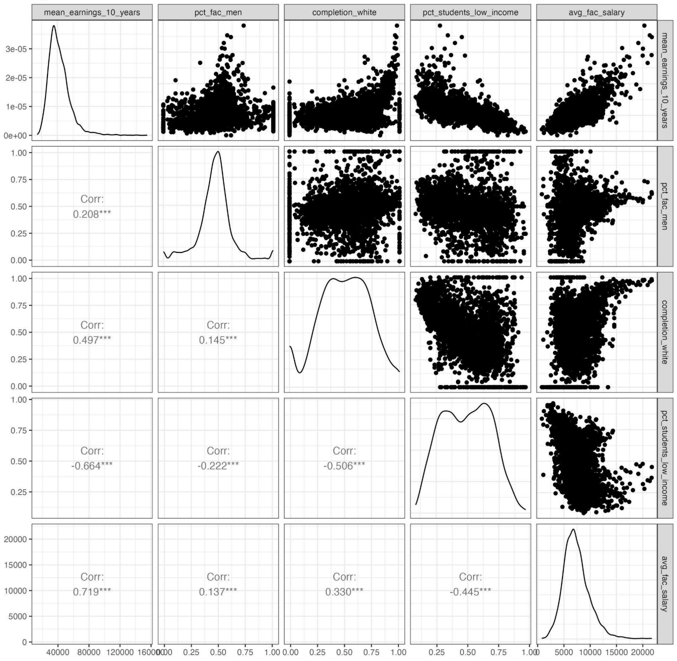


Figure 2. Histogram of mean earnings 10 years post-graduation in dollar

Figure 3. Pairwise Scatterplots between Quantitative Variables



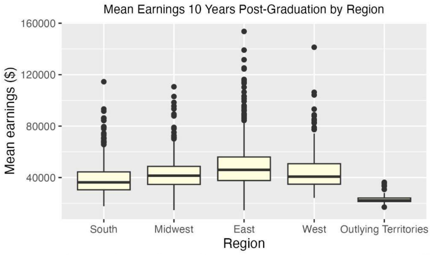


Figure 4. Boxplot showing the relationship between mean earnings 10 years post-graduation in dollars and region of institution

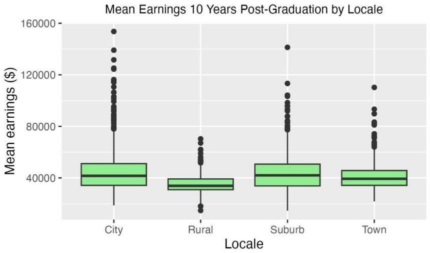


Figure 5. Boxplot showing the relationship between mean earnings 10 years post-graduation in dollars and locale of institution

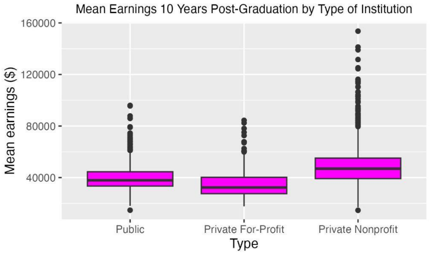


Figure 6. Boxplot showing the relationship between mean earnings 10 years post-graduation in dollars and type of institution

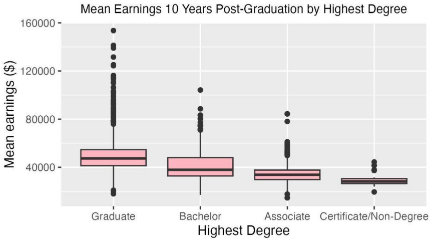


Figure 7. Boxplot showing the relationship between mean earnings 10 years post-graduation in dollars and highest degree offered at institution

Figures 2-7 show univariate and bivariate analyses for all the variables included in the final model (Figure 1 for the interaction is included in the main report). The histogram in Figure 2 shows that there are some potential high outliers in our response variable, which is otherwise approximately Normally distributed. The pairwise plots in Figure 3 don’t indicate too much concern about potential multicollinearity, and the boxplots in Figures 4-7 for the categorical variables indicate that there does seem to be some difference between the mean earnings of each category for each variable—though the concern about small sample sizes comes up again when we look particularly at the Outlying Territories category in the region variable and Certificate/Non-Degree category (which is already a combination of two categories with small sample sizes) in the highest degree variable.

Appendix C - Diagnostics for models (1) and (2)

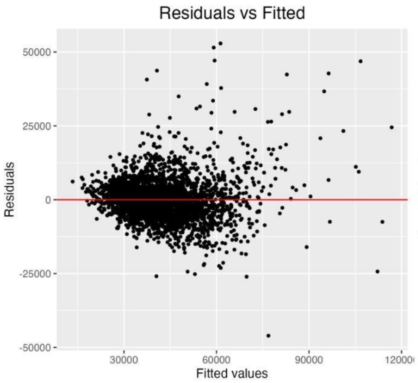


Figure 8. Residual plot for the multiple linear regression model (1)

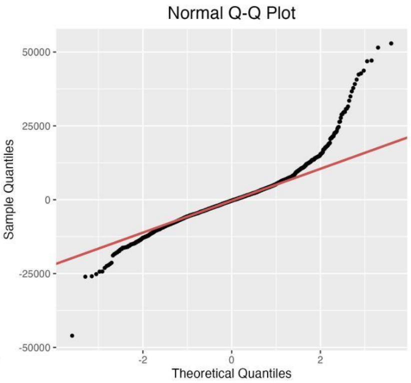


Figure 9. Normal probability plot for the multiple linear regression model (1)

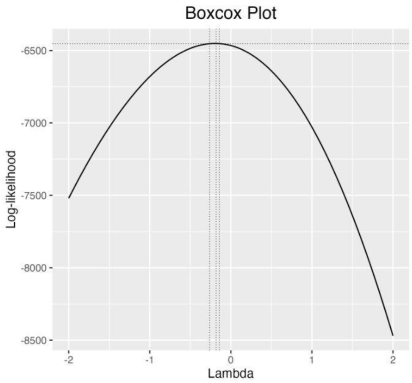


Figure 10. Boxcox plot for the multiple linear regression model (1)

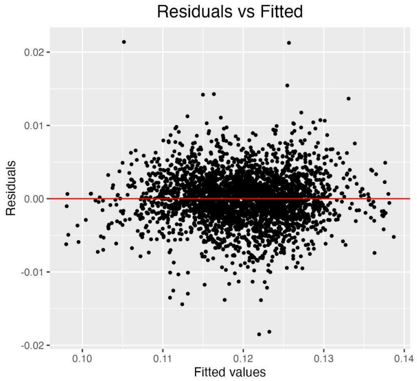


Figure 11. Residual plot for the multiple linear regression model (3)

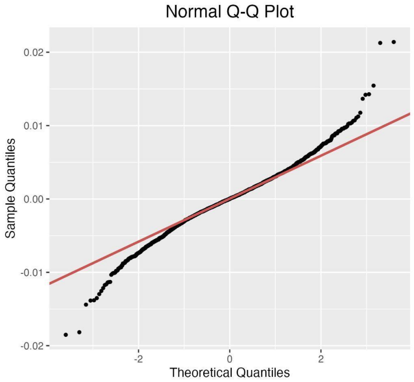


Figure 12. Normal probability plot for the multiple linear regression model (3)

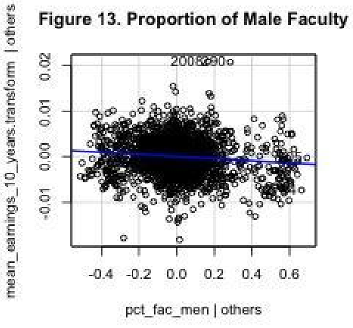


Figure 13. Proportion of Male Faculty

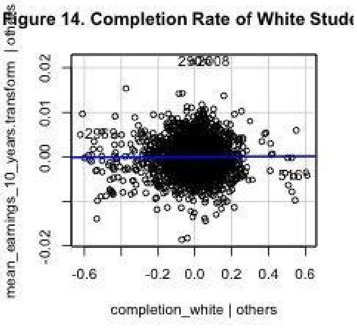


Figure 14. Completion Rate of White Stude

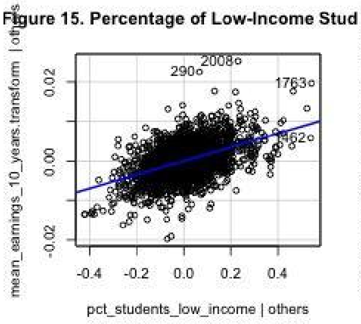


Figure 15. Percentage of Low-Income Stud

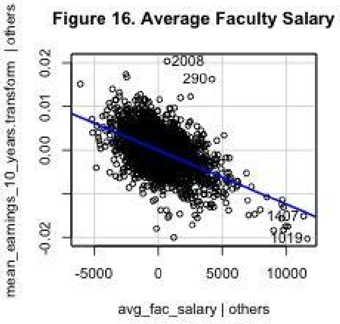


Figure 16. Average Faculty Salary

Appendix D - Regression Summary and ANOVA Table for Model (2)

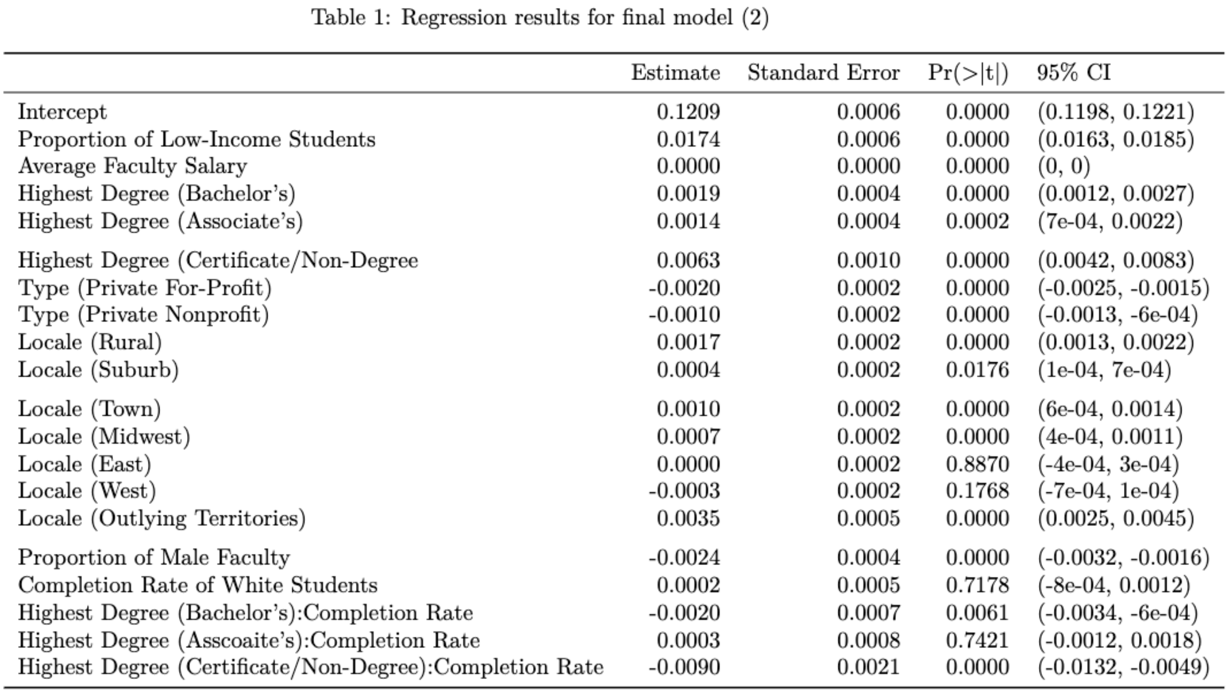


Table 1: Regression results for final model (2)

	Estimate	Standard Error	Pr(>|t|)	95% CI
Intercept	0.1209	0.0006	0.0000	(0.1198, 0.1221)
Proportion of Low-Income Students	0.0174	0.0006	0.0000	(0.0163, 0.0185)
Average Faculty Salary	0.0000	0.0000	0.0000	(0, 0)
Highest Degree (Bachelor's)	0.0019	0.0004	0.0000	(0.0012, 0.0027)
Highest Degree (Associate's)	0.0014	0.0004	0.0002	(7e-04, 0.0022)
Highest Degree (Certificate/Non-Degree	0.0063	0.0010	0.0000	(0.0042, 0.0083)
Type (Private For-Profit)	-0.0020	0.0002	0.0000	(-0.0025, -0.0015)
Type (Private Nonprofit)	-0.0010	0.0002	0.0000	(-0.0013, -6e-04)
Locale (Rural)	0.0017	0.0002	0.0000	(0.0013, 0.0022)
Locale (Suburb)	0.0004	0.0002	0.0176	(1e-04, 7e-04)
Locale (Town)	0.0010	0.0002	0.0000	(6e-04, 0.0014)
Locale (Midwest)	0.0007	0.0002	0.0000	(4e-04, 0.0011)
Locale (East)	0.0000	0.0002	0.8870	(-4e-04, 3e-04)
Locale (West)	-0.0003	0.0002	0.1768	(-7e-04, 1e-04)
Locale (Outlying Territories)	0.0035	0.0005	0.0000	(0.0025, 0.0045)
Proportion of Male Faculty	-0.0024	0.0004	0.0000	(-0.0032, -0.0016)
Completion Rate of White Students	0.0002	0.0005	0.7178	(-8e-04, 0.0012)
Highest Degree (Bachelor's):Completion Rate	-0.0020	0.0007	0.0061	(-0.0034, -6e-04)
Highest Degree (Asscoaite's):Completion Rate	0.0003	0.0008	0.7421	(-0.0012, 0.0018)
Highest Degree (Certificate/Non-Degree):Completion Rate	-0.0090	0.0021	0.0000	(-0.0132, -0.0049)

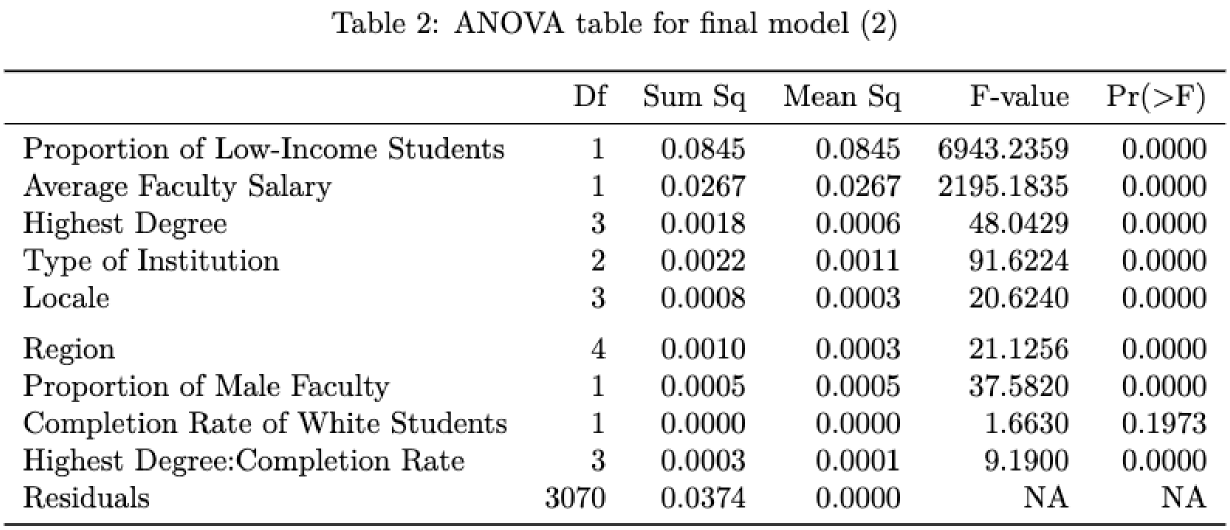


Table 2: ANOVA table for final model (2)

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Proportion of Low-Income Students	1	0.0845	0.0845	6943.2359	0.0000
Average Faculty Salary	1	0.0267	0.0267	2195.1835	0.0000
Highest Degree	3	0.0018	0.0006	48.0429	0.0000
Type of Institution	2	0.0022	0.0011	91.6224	0.0000
Locale	3	0.0008	0.0003	20.6240	0.0000
Region	4	0.0010	0.0003	21.1256	0.0000
Proportion of Male Faculty	1	0.0005	0.0005	37.5820	0.0000
Completion Rate of White Students	1	0.0000	0.0000	1.6630	0.1973
Highest Degree:Completion Rate	3	0.0003	0.0001	9.1900	0.0000
Residuals	3070	0.0374	0.0000	NA	NA