Outcome Gaps in Higher Education

Xiaoqian Zhu, Shirley Jin, Tina Huang, Abigail Chaver

December 5, 2016

1 Abstract

In this report, we investigate "outcome gaps" in higher education - the widespread phenomenon that underrepresented students have worse outcomes than their well-represented peers. Using data from College Scorecard, a federally maintained dataset, we explore the differences between completion rates across races, and differences between post-graduation earnings across parent-income terciles.

We test the hypotheses that the type of school (public, non-profit private, or for-profit private) and the instructional expenditure per full-time student are related to outcome gaps, finding statistically significant relationships in both cases. We also run a random forest regression across 14 other possible explanatory factors to provide some insight on possible interactions. The random forest model was not particularly useful as a predictor, but provided some interesting suggestions as to what circumstances correspond to worse outcome gaps.

2 Introduction

We are interested in factors that are related to "outcome gaps" for students in higher education. Our high-level goal is to identify factors which are associated with under-represented students having a higher degree of difficulty succeeding than their well-represented peers. We are investigating this issues partly from the perspective as students attending a public school, UC Berkeley, with great resources compared to other public schools, but less compared to private schools at similar levels of selectivity. We are attempting to provide a data-based analysis of the theory that underrepresented students struggle more in comparison to their peers when attending institutions that have fewer resources to support them.

Given that there is a major social-good imperative to decrease these outcome gaps, we hope that this analysis can provide hard evidence that there are patterns to the degree of outcome inequality. Specifically, we are interested in the question of whether additional government investment in public education can decrease outcome gaps. Obviously, it is impossible to conduct an experiment to answer this question and justify any kind of causal relationship, but with thorough investigation we hope to provide strong evidence that this pattern is consistent enough to be taken seriously by policymakers.

Since expenditures are also highly correlated with other factors - for example, there is no accounting for the higher cost of living in coastal and urban regions - it is worthwhile to consider some additional factors to see whether interactions between factors are related to outcome gaps. Running a random forest model will allow us to get some sense of which variables are most important for predicting outcome gaps. This broader view of what might be related to outcome gaps should provide more insight into how we can address the problem.

3 Data

3.1 Source

We began by looking at the available data from College Scorecard, available at `https://collegescorecard.ed.gov/`, which contains data on over 7500 higher education institutions in the US. College Scorecard has an API that can handle complicated filtered requests, but we downloaded the full dataset to have full flexibility

in exploration. College Scorecard provides a data dictionary to comb through the thousands of variables it provides.

3.2 Target Variables

We found that completion rates were disaggregated by race, and post-graduation earnings were disaggregated by parent-income terciles. Therefore, we computed outcome gaps for these two statistics according to the cohorts provided. We compared white completion rates to black, hispanic, and Asian completion rates, defining the outcome gap with the equation

$$[(WhiteCompletionRate) - (MinorityCompletionRate)]/(WhiteCompletionRate)$$

For earnings gaps, we have ordered terciles, so we compute

$$[(HigherTercileEarnings) - (LowerTercileEarnings)]/(HigherTercileEarnings)$$

While computing these outcome statistics, we found inevitably that some were incomputable. This was particularly common with historically black schools, where the completion rate for white students was not recorded or recorded as zero. Therefore, we excluded such observations from our data set, as we are interested in cases where we can compare outcomes between white students and students of color.

3.3 Explanatory Variables

We found that College Scorecard provided two explanatory variables that were useful for our analysis, "CONTROL", a coded variable representing whether the school is under control of the government, a corporation, or a non-profit. The dataset also provided 'INEXPFTE', Instructional Expenditures per Full-Time Student. This is an excellent proxy variable for a quantifier of the resources a school provides to its students.

In the third part of our analysis where we consider additional exploratory variables and their interactions, we used our best judgment to choose possible relevant variables. Our random forest regression uses variables related to the type of degree awarded, the region and type of location, the admission rate, the proportion of students of each of the four races considered, the median cost to attend, the median household income, and the poverty rate.

4 Exploratory Data Analysis

Our final dataset, after omitting null values, was around 2000 observations. We wanted to begin our informal analysis by plotting some data from schools with which are familiar. Therefore, we considered three cohorts - the California State University schools, the University of California schools, and the elite private schools, comprised of the Ivies and a few others. We actually considered a fourth cohort, the top-expenditure schools, but those mostly overlapped with the third cohort.

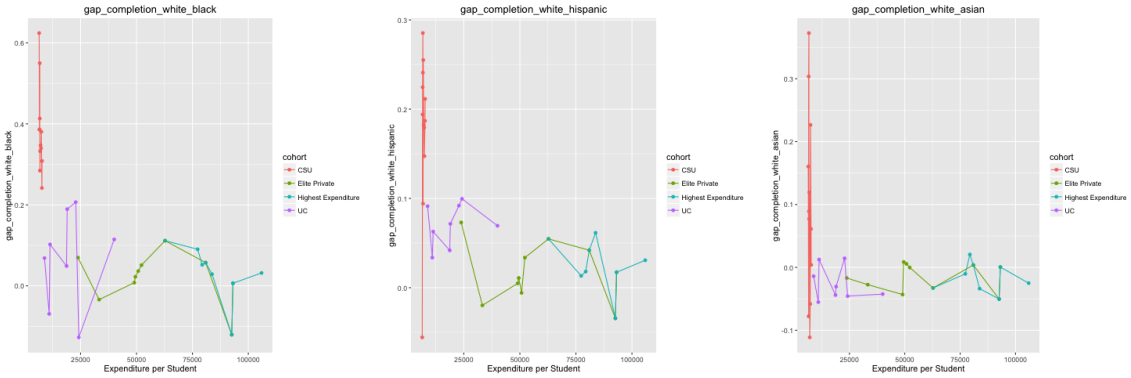


Figure 1: We see that while there is a clear segmentation of expenditures across cohorts, the difference in outcomes is not as clear-cut. There is a general negative relationship, but a lot of variance. The most obvious conclusion is that the CSUs suffer from much worse outcome gaps than the other cohorts.

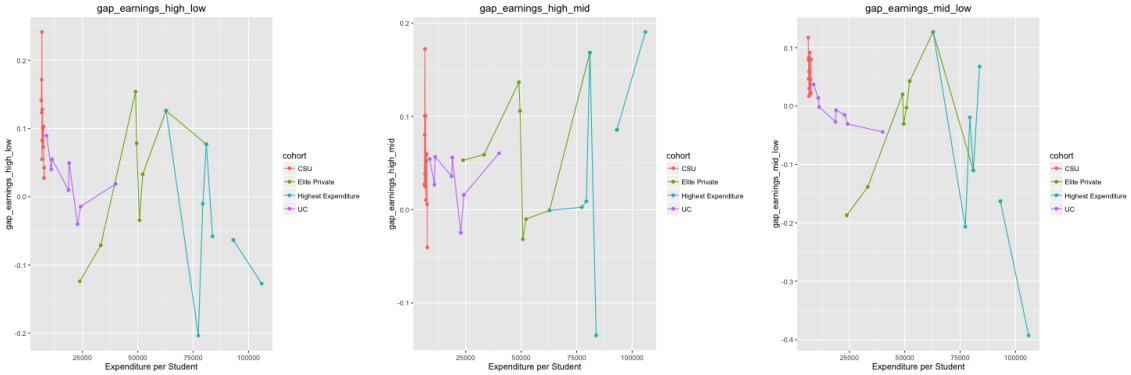


Figure 2: We see a similar pattern with earnings. Interestingly, the data actually suggests that middle-income students do worse in comparison to high-income students as expenditures increase.

We now look at the full dataset to get a sense of the relationship across all schools in the US. We begin with completion rates.

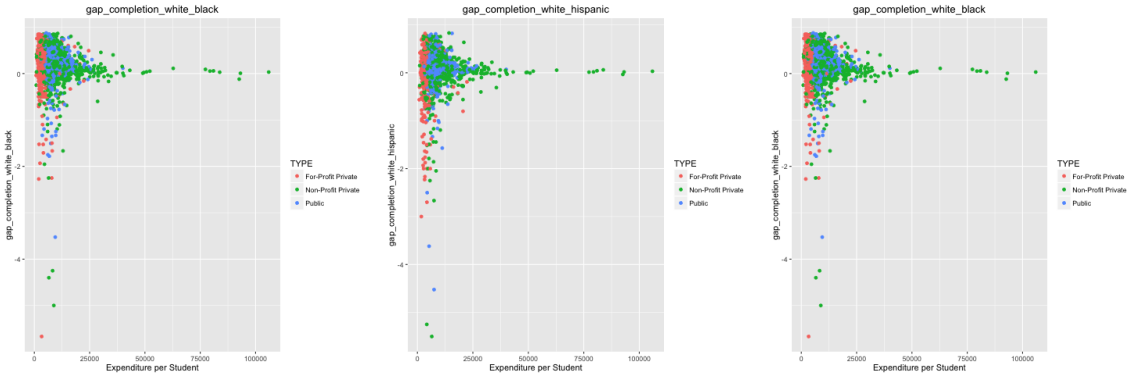


Figure 3: While there are definite differences in expenditure by type of school, it is not obvious that achievement gaps decrease with expenditures. Generally, achievement gaps are not centered too far from zero, and there are quite a few outliers in both dimensions. However, it does look like variance decreases and outcome gaps cluster more closely around zero as expenditures increase.

We restrict our bounds to exclude the outliers to see if we can visually detect any patterns.

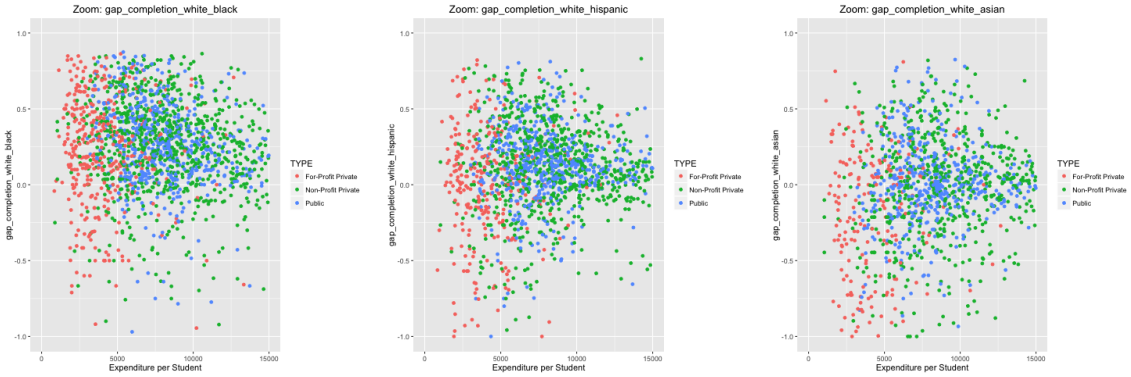


Figure 4: The data actually looks almost perfectly random.

Looking at our earnings metrics, we see a similar spread.

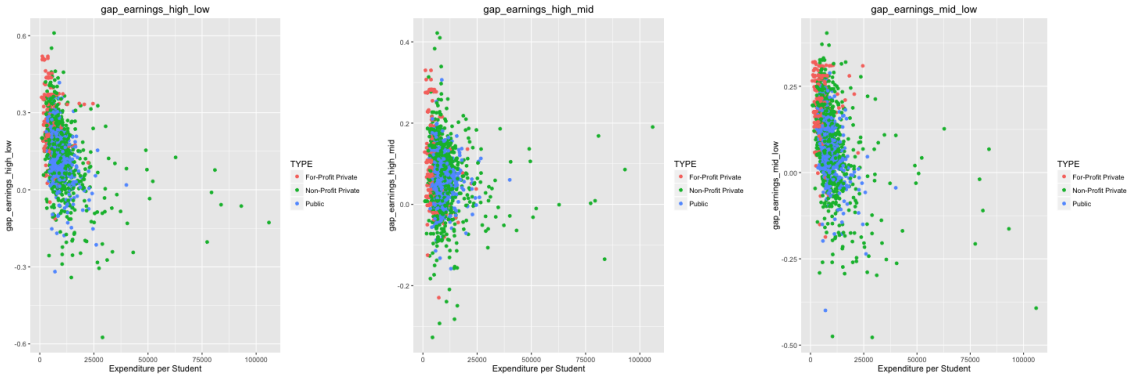


Figure 5: We see a similar pattern, where variance decreases and the gap clusters closer to zero as expenditure increases. An interesting observation is that the gap between medium and low income terciles actually seems centered below zero - low income students on average seem to perform a bit better than their middle income peers.

However, restricting bounds produces a more interesting visual insight.

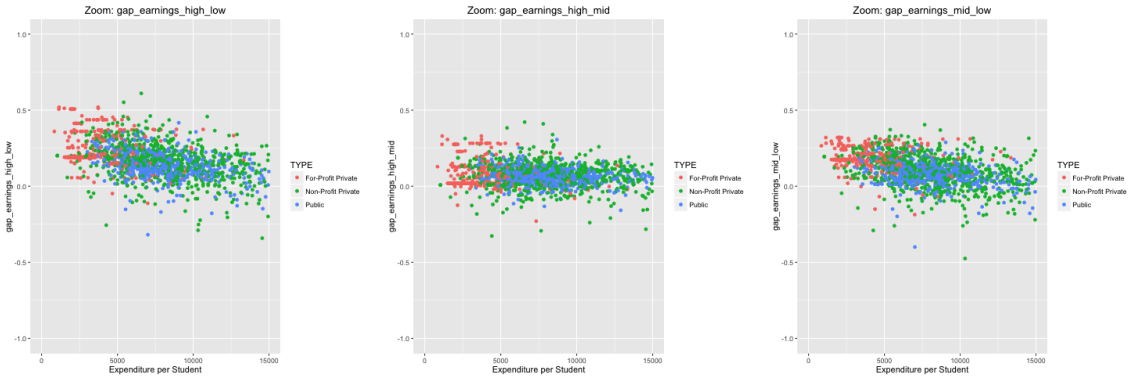


Figure 6: Fascinatingly, we actually see a very clear negative correlation between expenditure and outcome gap for High-Low and High-Medium. Medium-Low seems to have no relationship.

The relationship between expenditure and outcome gap is not obvious, but there definitely seem to be some non-random patterns worth investigating. In the next sections, we undertake formal statistical analysis to develop more concrete conclusions about these patterns.

5 Anova Analysis on Type of School

We want to first determine whether there is a significant outcome gap in terms of completion rate and earnings between the student groups at different types of institutions. Thus, we run an ANOVA model to assess whether there is a difference between public, private non-profit, and private school gap metrics. Let's first look at the analysis of variance of completion rates of black and white students.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cr_w_b$CONTROL	2	1.38	0.69	3.65	0.0261
Residuals	1938	366.44	0.19		

Table 1: ANOVA Table Completion Rate White and Black Students

As indicated by the high F statistics and low p-value, there is a significant difference between public, private non-profit, and private school black and white students completion rate gap metric. If we then look further for pair-wise comparisons in the Tukey Plot displayed, we see that the outcome gap is significant between Private For-Profit and Public schools, but not so much between Private For-Profit and Private Non-Profit or Private Non-Profit and Public schools.

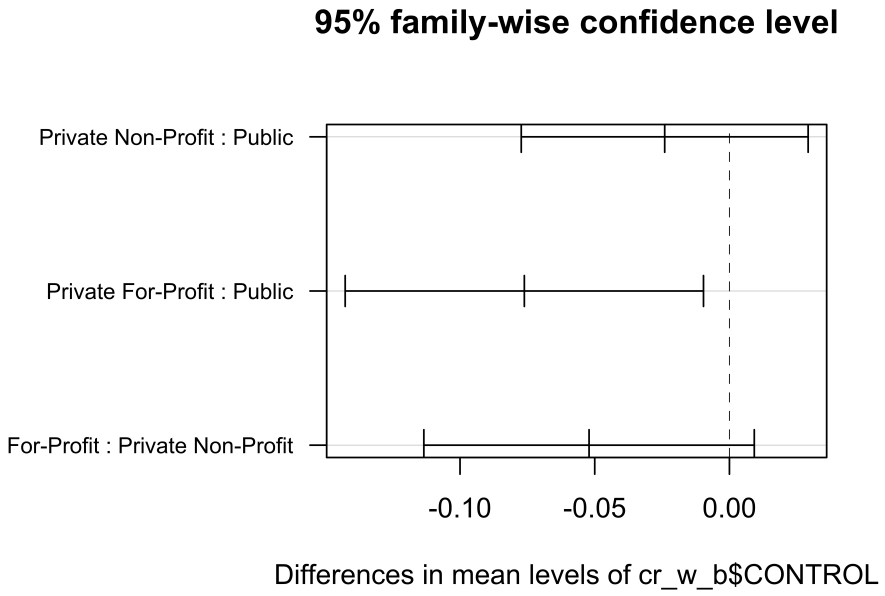


Figure 7: Tukey Plot Completion Rate White and Black Students

Doing similar analysis for completion rates by other racial groups, we see the completion rate gap metric is signifcant with Private For-Profit and Public schools in general. The gap metric is also signifcant with Private For-Profit and Private Non-Profit for Hispanic and White students.

Looking at the earnings gap metric amongst different student income groups, we see that the outcome gap is signficant amongst all institution types for low income students. The gap is significant between Private Non-Profit and Public schools for high and mid income students.

	Institutions	White_Black	White_Asian	White_Hispanic
1	Private Non-Profit w/ Public	0.54	0.63	0.82
2	Private For-Profit w/ Public	0.02	0.00	0.00
3	Private For-Profit w/ Private Non-Profit	0.11	0.00	0.00

Table 2: Completion Rate P-Value Table

	Institutions	High_Low	High_Mid	Mid_Low
1	Private Non-Profit w/ Public	0.00	0.20	0.87
2	Private For-Profit w/ Public	0.00	0.30	0.01
3	Private For-Profit w/ Private Non-Profit	0.00	0.01	0.00

Table 3: Earnings P-Value Table

6 Correlation Analysis on Expenditure per Student

We are trying to find the relationship between the gap metrics, including the completion gap and income gap, and school revenue per student, to see whether the school revenue per student has a negative correlation with the gap metrics.

6.1 Methodology

We start our analysis by setting up null and alternative hypothesis. The null hypothesis is, H_0, that there is no relationship between gap metrics and school expenditures per student; the alternative hypothesis is, H_1, that there is a relationship between gap metrics and school expenditure per student. These are equal to the following that $H_0 : \beta_1 = 0$ and $H_0 : \beta_1 \neg 0$. To test the hypothesis, we apply a simple regression model like $Gapterms = \beta_0 + \beta_1 INEXPFTE$. For both completion gap and income gap, we have three pairs of groups due to three race groups we have, thus we also have three regressions for each gap metric. In this case, we will use the least squares model on our data for the regression analysis.

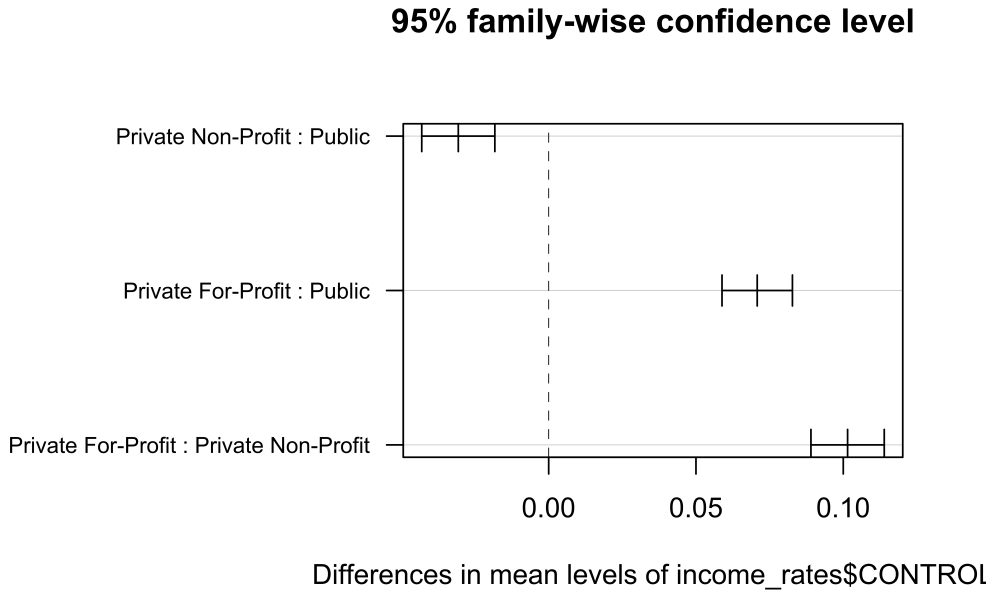
6.2 Results

Running regression through R, we can compute get the estimated coefficients. The regression coefficients for completion gap is given in the tables below:

Table 4: Completion Gap Slope comparison

	Model	Estimate	SE	p
1	Public: White-Black	-1.0446694E-05	3.8476020E-06	6.8215159E-03
2	Non-Profit Private: White-Black	-3.7046532E-06	1.4658166E-06	1.1608429E-02
3	For-Profit Private: White-Black	-5.1814583E-06	1.0226819E-05	6.1269699E-01
4	Public: White-Hispanic	6.5570747E-06	4.7176435E-06	1.6507761E-01
5	Non-Profit Private: White-Hispanic	2.7219058E-06	1.6211459E-06	9.3393803E-02
6	For-Profit Private: White-Hispanic	5.6599409E-06	1.2351433E-05	6.4706406E-01
7	Public: White-Asian	8.1540925E-06	5.9863531E-06	1.7372066E-01
8	Non-Profit Private: White-Asian	9.1393106E-06	1.8495301E-06	9.0738509E-07
9	For-Profit Private: White-Asian	3.5538606E-05	1.6410164E-05	3.1284125E-02

Many of our p-values are significant. However, we did not implement any methods to account for multiple testing. We visualize our coefficients for clarity:



H

Figure 8: Tukey Plot Earnings High and Low Income Students

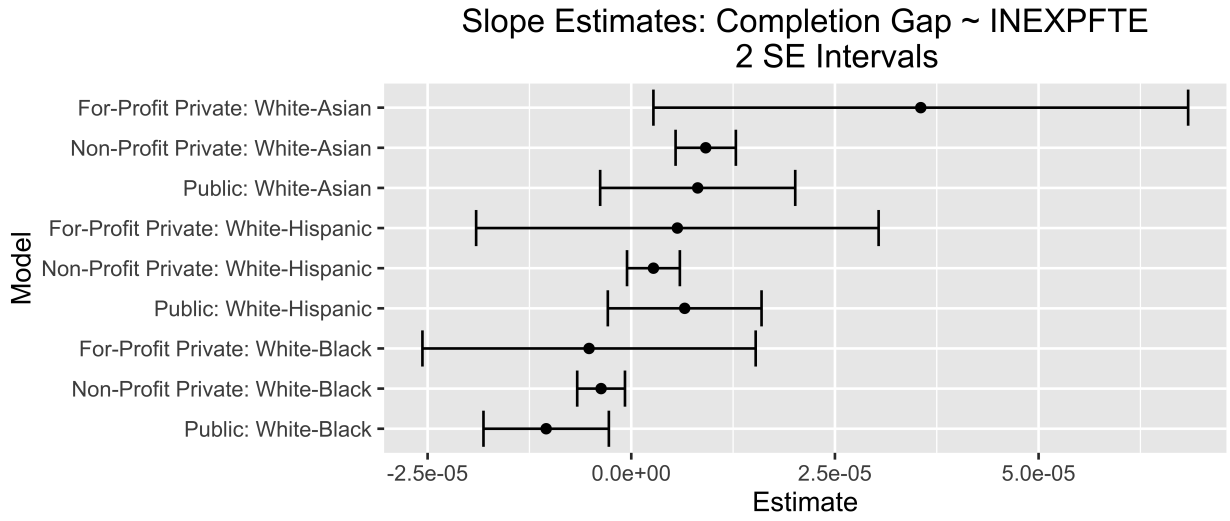


Figure 9: Generally, we are not left with an impression of overwhelming significance.

These p-values confirm our expectations from visual analysis, these coefficients are far more significant. We plot our coefficients again.

Table 5: Earning Gap Slope comparison

	Model	Estimate	SE	p
1	Public: High-Low	-9.2041360E-06	3.8399447E-07	5.1346929E-103
2	Non-Profit Private: High-Low	-5.7739524E-06	3.1945466E-07	8.5798680E-68
3	For-Profit Private: High-Low	-3.6563335E-06	9.6002048E-07	1.4781823E-04
4	Public: High-Mid	-2.5052138E-06	2.4407280E-07	1.0747897E-23
5	Non-Profit Private: High-Mid	-8.8340360E-07	2.3072470E-07	1.3269052E-04
6	For-Profit Private: High-Mid	-1.8751457E-06	8.4870800E-07	2.7360200E-02
7	Public: Mid-Low	-6.5401172E-06	3.1767348E-07	3.0206296E-80
8	Non-Profit Private: Mid-Low	-5.3375774E-06	2.6979047E-07	9.3261399E-80
9	For-Profit Private: Mid-Low	-2.5433119E-06	7.0384547E-07	3.1639444E-04

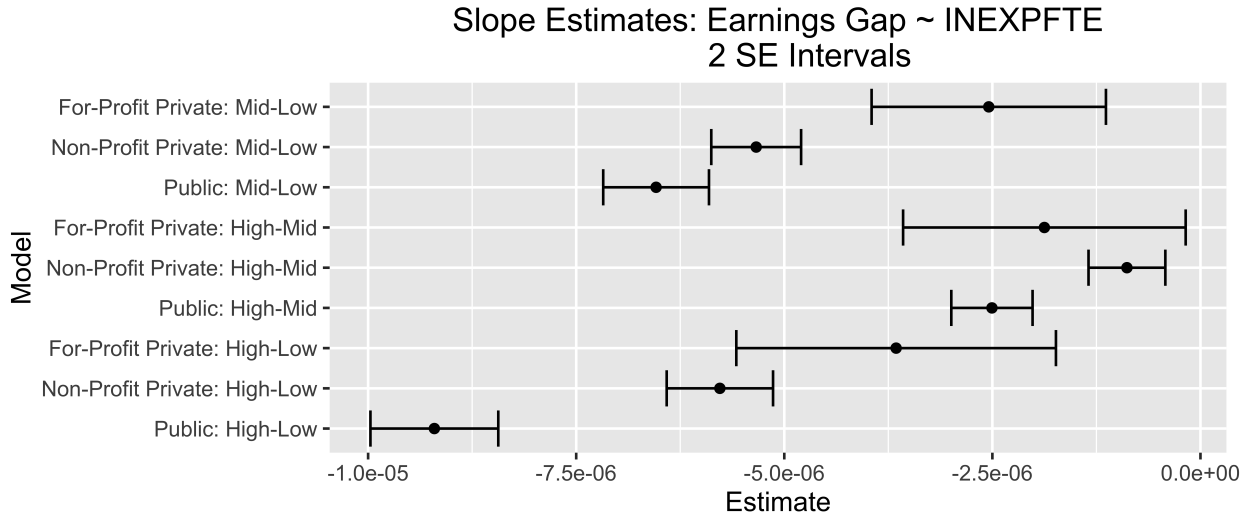


Figure 10: While again we should be concerned with multiple testing, we can be quite confident that there is a significant relationship between expenditures per student and earnings gap, especially between high and low income students. These coefficients also confirm the unexpected pattern that expenditures are positively correlated with low income students outperforming middle income students. However, we are less confident in these coefficients.

One interesting pattern in both graphs is that the standard error is much higher for for-profit private colleges.

Overall, we find ambiguous evidence that racial outcome gaps are correlated with expenditure, but we are much more confident in saying that there is a negative correlation with earnings gaps, especially the gap between high and low income students.

7 Random Forest

Our goal in this section is to do some inference on other possible predictors. We identified fourteen variables which we considered possibly relevant and briefly summarize them in Table 16.

Table 6: Description of Random Forest Explanatory Variables

	Explanatory.Variable	Description
1	SCH_DEG	Highest Degree Awarded
2	PREDDEG	Predominant Degree Awarded
3	REGION	Region of the US
4	LOCALE	Locale type (e.g. Urban = 12)
5	ADM_RATE	Admission Rate
6	UGDS_WHITE	White Proportion of Student Body
7	UGDS_BLACK	Black Proportion of Student Body
8	UGDS_HISP	Hispanic Proportion of Student Body
9	UGDS_ASIAN	Asian Proportion of Student Body
10	COSTT4_A	Average Cost to Attend
11	MEDIAN_HH_INC	Median Household Income
12	POVERTY_RATE	Poverty Rate of Students
13	INEXPFTE	Expenditures per Student
14	CONTROL	Institution Type

For each gap metric, we ran a random forest regression on the 14 variables, with 100 trees per forest. We checked the error plots to verify that we had reached a plateau in error at this number of trees. We also tried a few values for the number of variables to consider at each split, and got the best performance with 4.

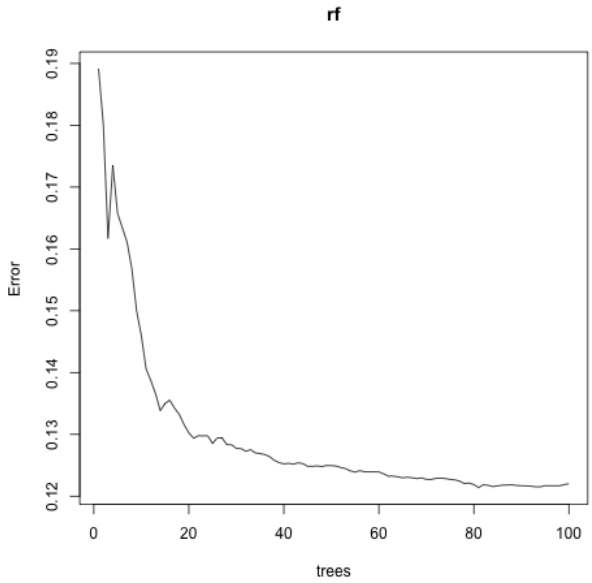


Figure 11: We see that increasing the number of trees is unlikely to produce better results.

In this situation, we are not attempting to build a predictive model. We are focused on identifying interesting relationships. Still, our models performed fairly poorly in terms of percent variance explained, so we should take these results with a grain of salt.

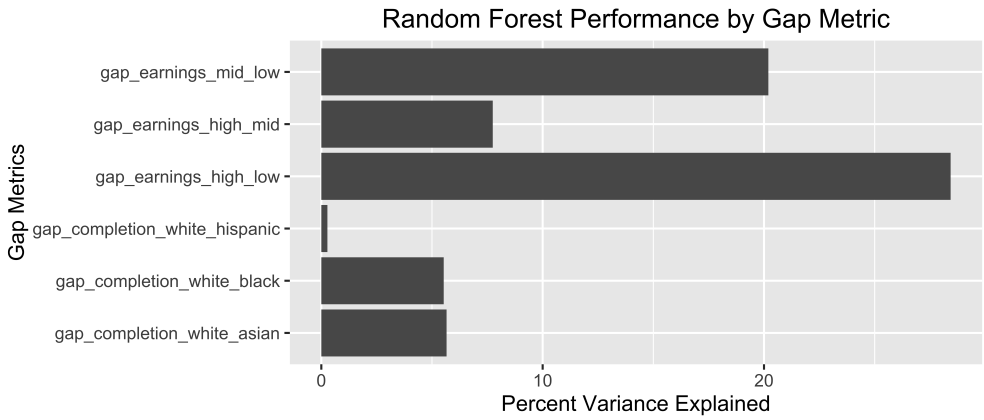


Figure 12: Most of our Random Forests performed extremely poorly. The best model was that which explained the variance between High and Low Income terciles. Interestingly, this recalls that the most visually noticeable trend in EDA was between these groups.

The reason we chose random forest was that this model provides a heuristic for variable importance. This statistic is the result of caculating the average increase in node purity associated with each variable over all one hundred trees. This metric is difficult to understand without some conception of how a decision tree is constructed, but an abstract explanation is that the statistic indicates how much more accurate a tree is in predicting the target when given the variable. Therefore, a comparison of node purity increase across the variables for the separate models can provide some insight for our purposes.

Because the spread of the completion rate gaps was much larger than the spread of earning gaps, we consider these types separately.

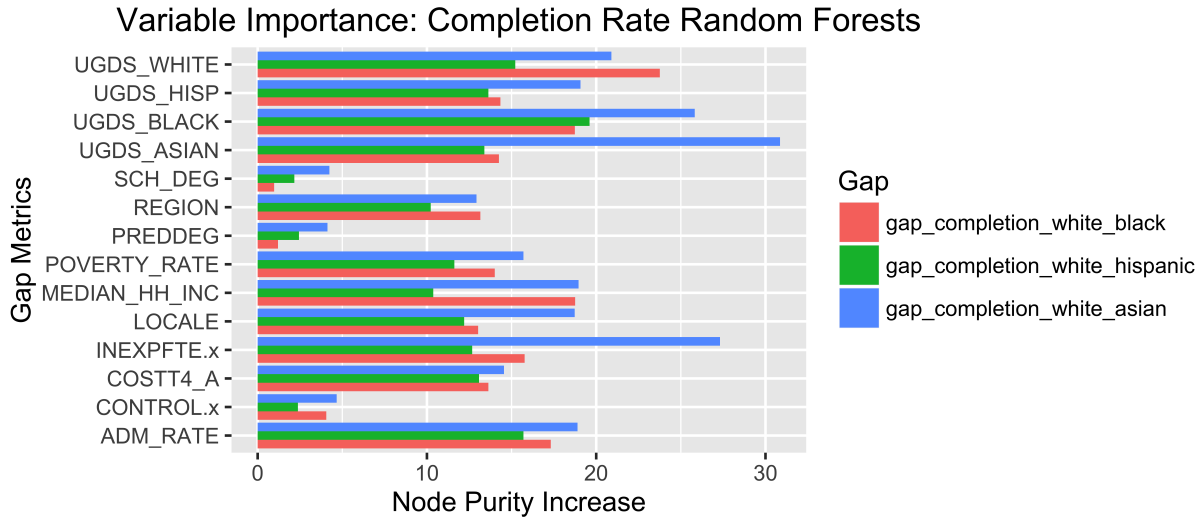


Figure 13: The most notable takeaway is that INEXPFTE is in fact one of the most important predictors, especially for the white Asian gap. Recall that this model performed the best of the completion rate models. Another noticeable trend is that the proportion of the student body which is Asian is a strong predictor for the white Asian completion gap. We also see a similar trend with the UGDSBlack for the white black completion gap. Generally, the diversity of the student body seems to be fairly relevant.

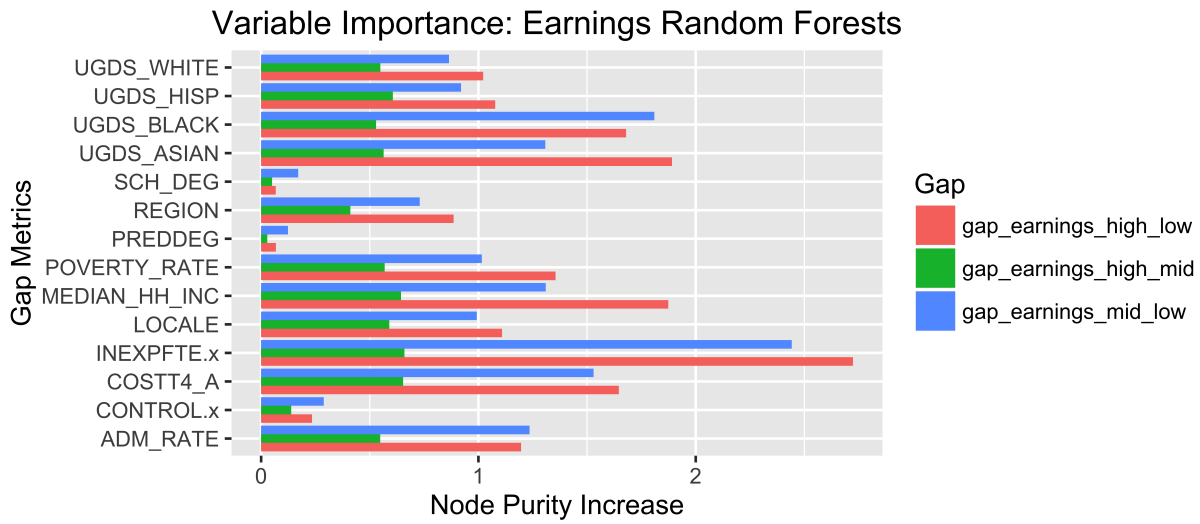


Figure 14: Here again, we see that INEXPFTE is the strongest variable for the best performing model, high low. We also notice that Median Household Income and Cost to Attend are stronger in these models, which intutitively makes sense given that we are comparing groups according to parent income. Interestingly, UGDSBLACK and UGDSASIAN are fairly important as well.

Relating these results to our previous analysis, we find that overall, Expenditures per Student seems to have a stronger relationship with outcome gaps than most other variables, while CONTROL is actually consistently one of the weakest variables.

8 Conclusion

In this report, we investigate outcome gaps in higher education with a particular focus on the relationship with school expenditures. While the significance of our results varied, we did find some significant patterns confirming a relationship. Our most notable conclusion is that there is a strong negative linear relationship between school expenditures per student and the post-graduation earnings gap between students of high-income backgrounds and low-income backgrouds. We see visually that at around $15,000 per year, the income gaps reach their ideal point, which is clustered evenly around zero.

It is somewhat irresponsible to draw causal conclusions from this observed pattern. However, it is notable that compared to many other variables, instructional expenditure per student stands out as the most important variable of many in our random forest models. While school expenditure cannot tell the whole story, there is compelling evidence that there is a relationship between the resources devoted to education and reduced educational inequality.

One counterargument is that underrepresented students who are better prepared for college attend better schools, which tend to have more resources to spend on students. This is entirely possible. However, we would point out that in the random forest analysis, Admission Rate, a good proxy for school prestige, was consistently a less important factor than instructional expenditure and most of the statistics regarding student body racial diversity. Additionally, for the random forests cohorted by income, Median Household Income is a very important variable. As we are not being too shy about our political perspective, we note that this is consistent with the theory that the outcomes of underrepresented students are affected by a lack of racial and economic diversity, and by the amount of resources that their school can offer them.

While we are approaching this topic with something of an agenda, we have been careful to acknowledge results which are insignificant or inconsistent with our hypothesis. Overall, though, we find that there is evidence supporting our hypothesis that per-student expenditure is related to outcome gaps. The less significant difference between types of schools suggests that school organization is less important than resources, which should be relevant to those interested in decreasing educational inequality.