An Analysis of college type and its net costs

Ruizhe Huang - 1006444331

December 17, 2021

Abstract

This report aims to give some guidelines regarding the college types and its net price for those who want to apply for a college in America. I hypothesize that private colleges may have higher net costs than public colleges. The dataset is about the top 650 American colleges ranking in 2019. The dataset is from Kaggle, and I cleaned the original dataset. The variables description and numerical and graphical summaries can be found in the data section. Several methodologies are used in this report, including the propensity score matching, two-sample t-test, logistic regression model and multiple linear regression model. The result section shows that the population mean of the net price for public and private colleges is different; that is, the net price varies between public and private colleges; however, it may result from the confounding variables within the dataset. Therefore, I perform the propensity score matching method to reduce or eliminate selection bias in observational studies. After all, we will know that there will be a causal inference between college type and net cost. In conclusion, since the college is private, on average, it has a higher net price than the public college. More specifically, on average higher than $13084.9 per year. Moreover, there are more private colleges than public colleges in America. More specifically, private colleges are around 1.5 times of public colleges in America.

Keywords: propensity score matching, causal inference, two-sample t-test, logistic regression, multiple linear regression, American colleges data

Introduction

The cost of college today is much higher than it was decades ago. Paying for college is expensive, and many families have felt the pinch of rising tuition over time. "Typically, private liberal arts colleges have the highest sticker price, but that does not mean they will not be competitive with other universities given their large endowments – especially if you qualify for financial aid and scholarship opportunities," says Amy Goodman Miller, master college admissions counsellor at IvyWise, a New York-based admissions consulting company. (Powell et al., 2021)

I hypothesize that even though subtracting the financial aid, on average, private colleges still have higher net costs than public colleges as private colleges rely on student tuition fees, alumni, and endowments to fund their academic programs.

I got the dataset for this analysis regarding the top 650 American colleges ranking in 2019; colleges are ranked based on alumni salary (20%), student satisfaction (20%), debt (20%), American leaders (15%), on-time graduation rate (12.5%), and academic success (12.5%).(Conklin, Coudriet,& Howard, 2019)

The research question is: **What is the relationship between the college type and its net costs?**. It is important to give some guidelines for those who want to apply for a college as the choice of colleges could be a profound decision and even could lead to the family financial burden.

I used the methodologies introduced in the method section to investigate the relationship between the college type and its costs. I will create a multiple linear regression model that uses the college's net price as the response variable. Moreover, I will use the propensity score matching method to mitigate the impact of confounding variables that may also affect the net costs of the college.

In the data section, there will be a data description in detail. Moreover, I have numerical summaries and graphical summaries of net price and Public/Private.

In the methods section, I will introduce the two-sample t-test to test whether the population mean of the average net price for public college is the same as that of the private college and estimate the propensity score using a logistic regression model were treatment (college types) is outcome based on the vector of covariates(student population, alumni salary and acceptance rate). Eventually, use the net costs as the response variable of the multiple linear regression model to evaluate the outcome of the propensity score matching.

In the results section, the net price varies between public and private colleges, and we get the causal inference of college types and the net costs.

In the end, there will be a conclusion to summarize the main points of this report.

Terminology

Net Price: The variable of interest is the American college's average cost per year of education, subtracting any financial aid received by the students.

Rank: The rankings are based on six general categories: Alumni Salary (20%), a combination of early and mid-career salaries as reported by the federal College Scorecard and PayScale data and research; Student Satisfaction (20%), which includes results from Niche surveys on professor quality and data, and freshman retention rates from the federal IPEDS website; Debt (20%), which rewards schools for low student debt loads and default rates; American Leaders (15%), which is based on the Forbes database of successful people, including billionaires, powerful women, 30 Under 30 honorees, leaders in public service and in private enterprise, and more; On-Time Graduation Rate (12.5%), which accounts for both four- and six-year rates; and Academic Success (12.5%), which rewards schools whose alumni win prestigious scholarships and fellowships like the Rhodes and the Fulbright or have earned Ph.Ds.(Conklin, Coudriet,& Howard, 2019)

logistic regression model: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is a predictive analysis. Logistic regression does not make many of the key assumptions of linear regression models based on ordinary least squares algorithms. Particularly regarding linearity, normality, homoscedasticity, and measurement level (Assumptions of logistic regression, 2021).

Propensity score matching: It was created for participants in the treatment and control groups. A matched set consisted of at least one participant in the treatment group and one in the control group

with similar propensity scores. The goal is to approximate a random experiment, eliminating many of the problems associated with the analysis of observational data (Stephanie, 2021).

Data

Data Collection Process

The dataset is collected from the Forbes magazine that select colleges and universities that educate undergraduates, doctoral research universities, master's universities and colleges and baccalaureate colleges. It also included colleges that provide specialized four-year engineering, business, and art programs. The data is from the following sources: The College Scorecard and the Integrated Postsecondary Education Data System (IPEDS), two federal databases that track student outcomes and institutional characteristics. It also drew from PayScale, a salary reporting and comparison company; Third Way, a D.C.-based think tank; and the National Center for Science and Engineering Statistics (NCSES). It cut colleges with fewer than 300 undergraduates (Kreznar, 2021). Therefore, this is an observational dataset without specific control on the variables.
However, the foreseeable limitation could be the data collected from 2019(pre-COVID-19 period). The net cost may differ after the COVID-19 pandemic. Moreover, the dataset only contains the top 650 colleges in America. Thus the data may not be a perfect representative sample.

Data Summary

The dataset is a CSV table regarding the top American colleges and the corresponding information. The dataset was directly downloaded from Kaggle (URL: https://www.kaggle.com/chris95cam/forbes-americas-top-colleges-2019). The source of this dataset is from Forbes Magazine's rankings of 650 U.S. colleges. The dataset contains the rankings of 650 Unites States colleges in 2019 along with various other statistics of each college such as student population, total annual cost, alumni salary, acceptance rate, etc.(Conklin, Coudriet,& Howard, 2019)
The original dataset includes 650 observations and 17 variables. Each observation represents a college in the America. The original dataset includes some missing values that may need to be cleaned.

Data cleaning process

The dataset was directly downloaded from Kaggle. It contains many missing values and variables that are not useful to my analysis.
I removed the variables that are not useful to my analysis, such as the *website* of the college. Moreover, I removed several variables that may have multicollinearity with the others. For example, the *Student Population* already includes the *Undergraduate Population*. The *Net Price* is calculated by the *Total Annual Cost* and *Average Grant Aid*. Therefore, only 7 important variables were kept: *Rank*, *Name*, *Public/Private*, *Student Population*, *Net Price*, *Alumni Salary* and *Acceptance Rate*. These variables are major factors to consider in high school students' college choices. For example, rankings provide a college's reputation. In this analysis, private or public colleges are the variable of interest to determine which type of college has a higher net price. We may know how much potential future alumni resources are by student population. Net price provides a reference for families to make financial plans. High school graduates can use alumni salaries to determine how much they can expect to earn 10 years after graduation. The acceptance rate is crucial to know the probability of college acceptance.
I renamed the variable names from coded names to the common language. I created another new dataset that converts the private college to 0 and the private college to 1 for later logistic regression analysis. I also removed all the missing values of the 7 important variables and provided a comparison numerical summary table in the later section before and after removing the missing values.
The cleaned dataset contains 632 observations and 7 variables without any missing value.

Variable Description

All the columns within the cleaned dataset are as following.

Table 1: Description of important variables

Variable	Description	Type
Rank	College ranking by Forbes Magazine	Numerical
Name	Name of the college	Categorical
Public/Private	Whether school is publicly or privately funded	Categorical
Student Population	Total number of students enrolled	Numerical
Net Price	Average cost for one year of education	Numerical
Alumni Salary	Median salary for the alumni	Numerical
Acceptance Rate	Percentage of students who apply to a college that are admitted	Numerical

Table 1 is a description of selected important variables. I kept only a few variables in the cleaned data to give readers a picture of the essential factors for each college. The ranking is based on alumni salary (20%), student satisfaction (20%), debt (20%), American leaders (15%), on-time graduation rate (12.5%), and academic success (12.5%).
In this report, I will only focus on the following variables *Public/Private*, *Student Population*, *Net Price*, *Alumni Salary* and *Acceptance Rate*.
Public colleges represent the educational institutions that are mainly funded by state governments. *Private* colleges rely on student tuition fees, alumni and endowments to fund their academic programs.
The *Student Population* is the total number of students enrolled in the college.
The *Net Price* is one of the variables of interest. The college's average cost per year of education is subtracting any financial aid received by the students.
The *Alumni Salary* is the median salary for workers with 10 or more years of experience count in US Dollars per year.
The *Acceptance Rate* is the percentage of students whom the colleges admit, that is, admitted students divided by the number of students who apply to the college.

Numerical summaries

Table 2: Numerical summaries of the original numerical variables

Variables	Mean	Median	Min	Max	IQR	Missing Values
Student Population	12022	6269	386	75044	15547	0
Net Price	22337	21989	0	47270	11170	2
Alumni Salary	98852	96400	70700	158200	34900	15
Acceptance Rate	61.6	67	5	100	30	2

Table 3: Numerical summaries of the cleaned numerical variables

Variables	Mean	Median	Min	Max	Standard Deviation	IQR	Missing Values
Student Population	12258	6488	406	75044	13223.21	15851	0
Net Price	22339	22012	0	47270	8268.86	11430	0
Alumni Salary	98867	96400	70700	158200	14366.97	17100	0
Acceptance Rate	61.4	67	5	100	22.09	30	0

Table 2 and 3 is the numerical summaries of the numerical variables before and after removing the missing values. As we can see, the numerical summaries of these variables are not varied a lot after removing the missing values. Therefore, we have reasonable ground to remove the missing values of the chosen variables. From table 3, among 632 colleges, the sample mean student population is 12258, and the sample median is 6488, lower than its sample mean. The net price has the sample mean of $22339, and the sample median is $22012, which is slightly lower than its sample mean. The alumni salary has the sample mean at $98867, and the sample median is $96400, which is slightly lower than its sample mean. The sample mean is 61.4%, lower than its sample median of 67% for the acceptance rate.
The sample standard deviation of net price is at $8268.86. The sample standard deviation of alumni salary is at $14366.97. It implies that the fluctuation of alumni salary is greater than the net price. The sample range of net price is from $0(United States Military Academy) to $47270(Berklee College of Music), and for alumni, salary is from $70700(Emory & Henry College) to $158200(Harvey Mudd College). It implies that the range of alumni salary is greater than the range of the college's net price. The sample IQR of the net price of college is less than alumni salary. The alumni salary range between the 1st and the 3rd quartile is larger. It implies that the spread of the college's net price is less than the alumni salary.
The sample standard deviation of the student population is 13223.21, which implies the student population fluctuate significantly between the colleges. The sample standard deviation of the acceptance rate is 22.09%, which implies the acceptance rate has a moderate fluctuation between the colleges. The sample range of the acceptance rate is quite extensive, from 5% (Stanford University) to 100% (University of Texas, El Paso). The sample IQR of acceptance rate is 30%, which means that the range between the 1st and the 3rd quartile is 30%.
There will be some graphical summaries analyzing them later as well.

Table 4: Proportion of the private or public college

term	Private college	Public college
proportion	60.76%	39.24%
count	384	248

From table 4, in the cleaned dataset, there are 384 private colleges, about 60.76%. There are 248 public colleges, about 39.24%. Private colleges account for the most among all the colleges in America. It is around 1.5 times of public colleges.
Furthermore, as we will perform the propensity score matching later, from table 4, we have 248 public colleges. Thus we expect the matching dataset contains 496 observations.

Graphical summaries

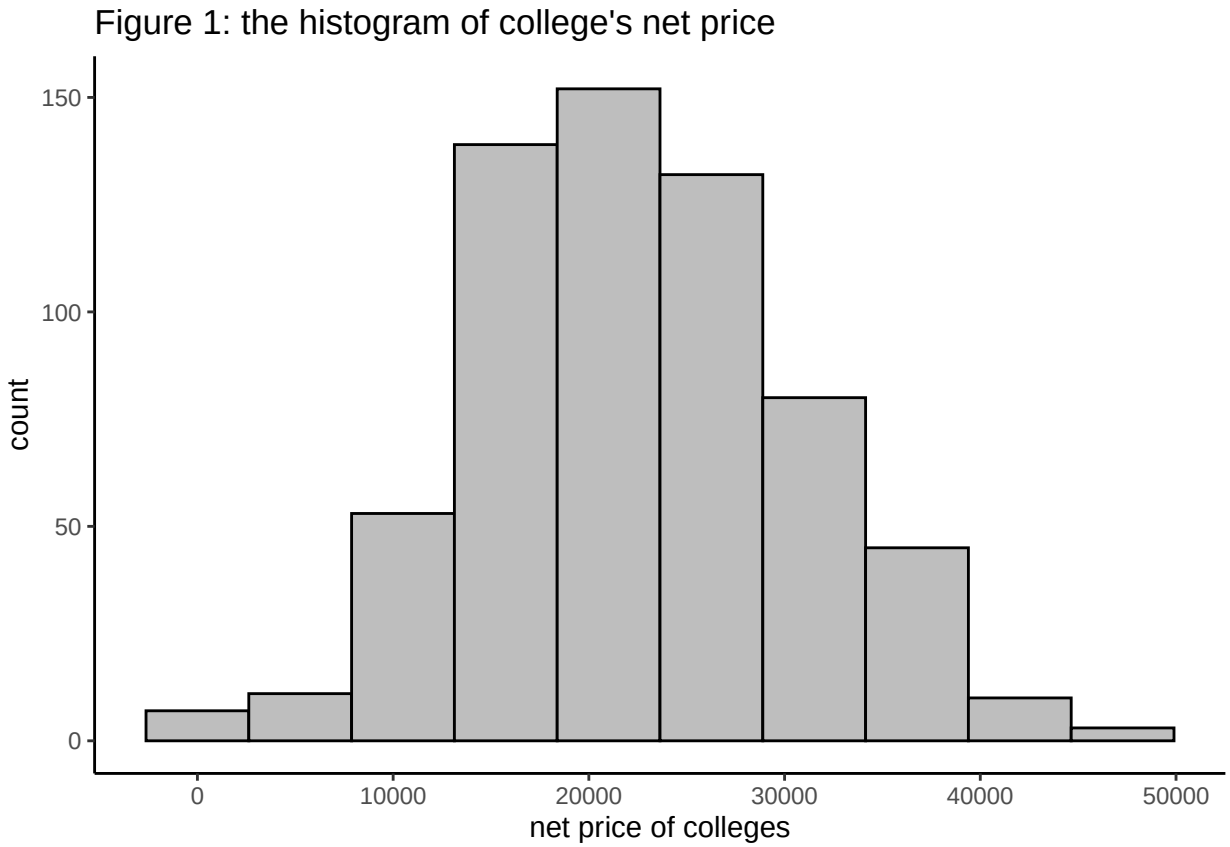

Figure 1 is the histogram of the net price of colleges. The shape of this histogram is almost symmetric, like a bell curve. It has a unimodal around $20,000. The centre of the histogram is around $20,000. It means that we expect to cost $20,000 per year in most colleges approximately. The spread of the college's net price is around $0 to $50,000 per year. Most of the time, we found that the college's net price is between $15,000 and $30,000 per year.

Since the net price of colleges is used as the response variable for the multiple linear regression model, thus, it is supposed to follow a Normal distribution, which is the assumption of the linear model. From figure 1, we can see that the distribution of net price of colleges is almost symmetric like a bell curve. It could give a very rough idea of the Normal assumption.

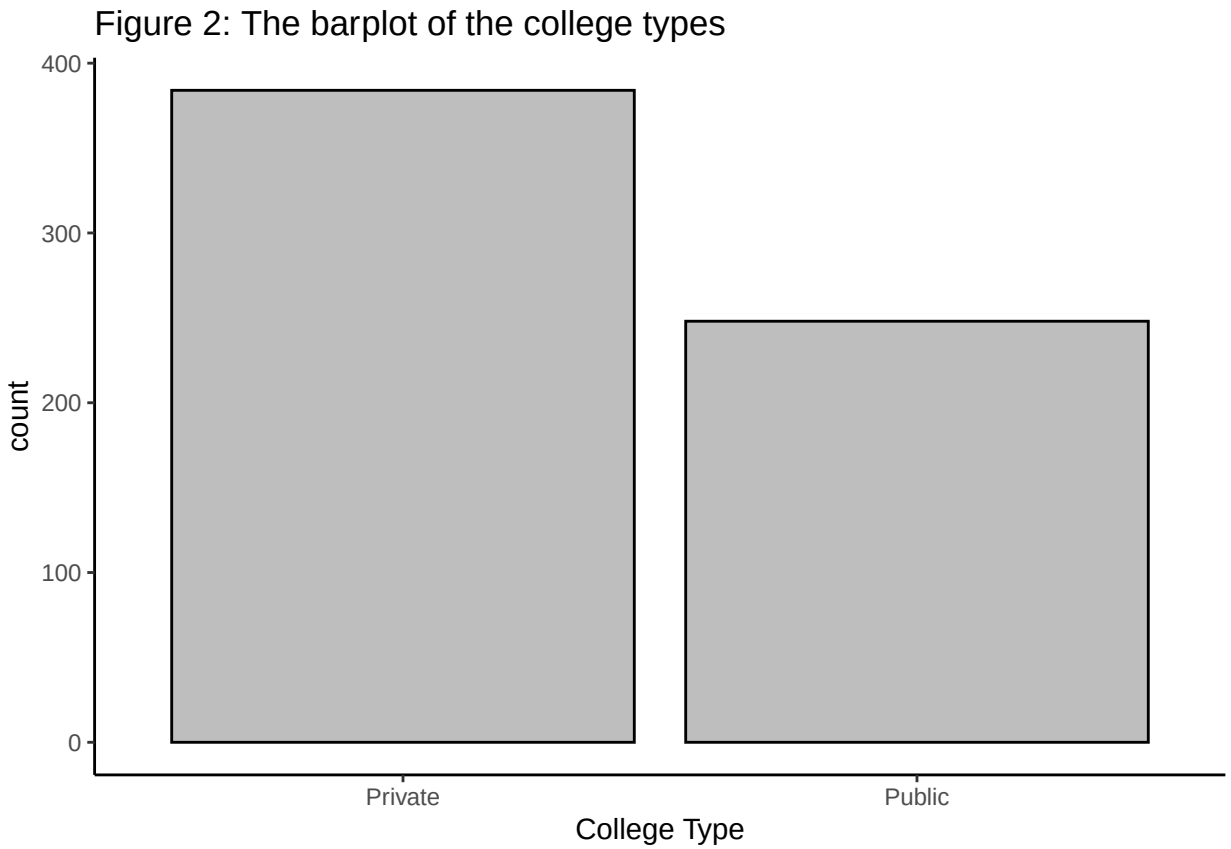

Figure 2 is a barplot of the college types(categorical variable). The private colleges accounted for the most significant proportion. The public colleges accounted for a relatively small proportion. The graph shows nearly 400 private colleges and around 250 public colleges. The number of private colleges is about 1.5 times that of public colleges. Hence by figure 2, we found that private colleges are more common than public colleges in America.

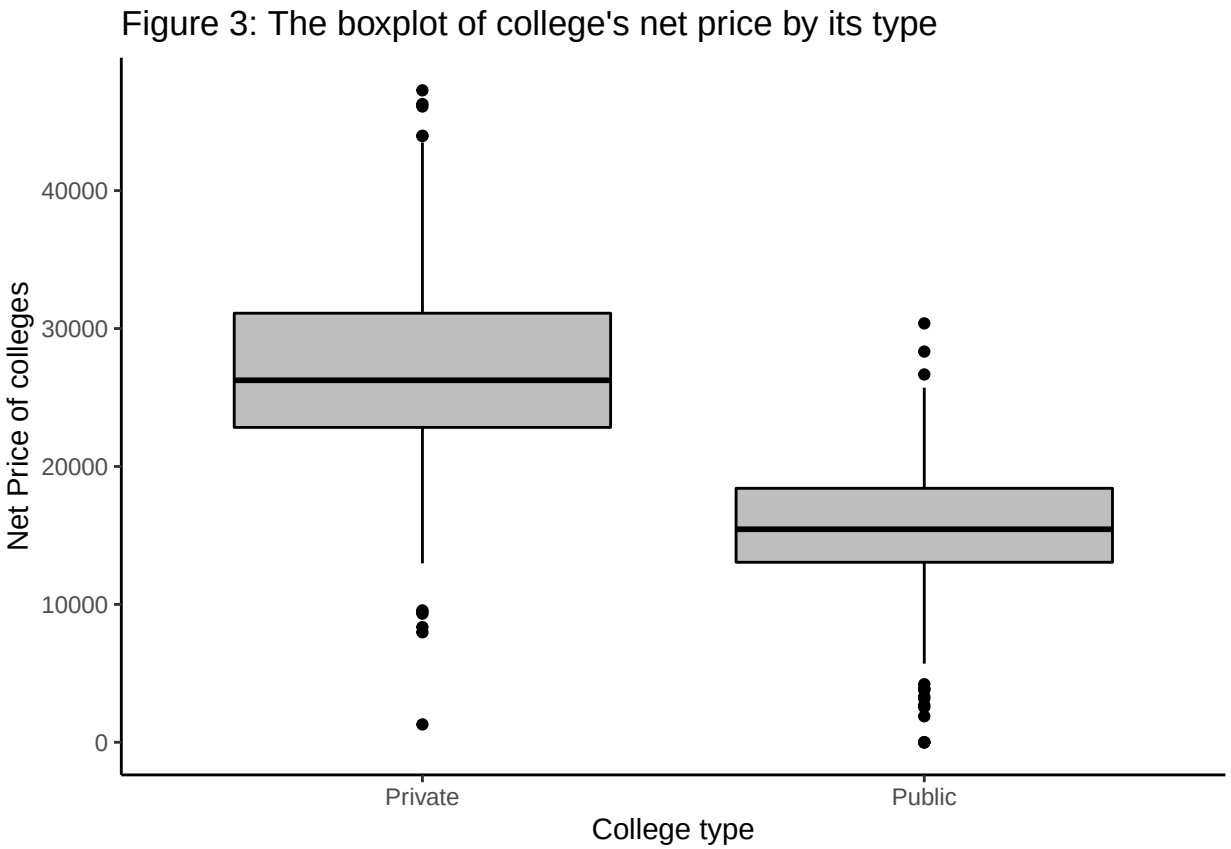

Figure 3 shows the boxplots of the college types. For the private colleges, the shape of the boxplot is slightly right-skewed with several outliers under the net price of $10,000 and above $40,000. The median net price of private college is centred around $25,000. The first quantile nearly at $22,000. And the third quantile nearly at $31,000. So the IQR of the net price of private college is approximately $9,000. It implies that the spread of the net price of private colleges is relatively large. In contrast, for the public colleges, the shape of the boxplot is almost symmetric with several outliers under the net price of $5,000 and above $25,000. The median net price of private college is centred around $15,000. The first quantile is around $12,000. Furthermore, the third quantile is around $18,000. So the IQR of the net price of private college is approximately $6,000. It implies that the spread of the net price of private colleges is relatively small.
In conclusion, the spread of the net price of private colleges is larger than the public college. The median net price of private colleges is much higher than public colleges. Therefore, we would expect a higher net price for private colleges and fluctuate more significantly than the public colleges.

Methods

To analyse the relationship between college type and its costs, I will first use the two-sample t-test to test whether the population mean of the average net price for public college is the same as that of the private college. If they are different, the net price varies between public and private colleges. However, it may result from the confounding variables within the dataset. Therefore, I will use the propensity score matching method to reduce or eliminate selection bias in observational studies.
Assume the data is independent, and the response variable, *Net Price* is normally distributed in the multiple regression model. Thus the sample mean is normally distributed as well.

Two sample t-test

The two-sample t-test is a method used to test whether the unknown population means of two groups are equal or not by the sample means of two groups.
The null hypothesis is $H_0 : \mu_{public} = \mu_{private}$, which means that the population mean of the average net price for public college is the same as the private college.
The alternative hypothesis is $H_A : \mu_{public} \neq \mu_{private}$, which means that the population mean of the average net price for public college is different from the private college.
Where μ_{public} is the population, mean of the average net price for public college, and $\mu_{private}$ is the population mean of the average net price for private college.
We will first use the r code to find out the test statistic. Assume the population mean of college's net price between public and private colleges are the same. Then simulate the difference of sample mean of college's net price between public and private colleges. Eventually, we will get its p-value, the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct (Beers, 2021). I will pre-specified the significance level $\alpha = 0.05$ as a cut-off for rejecting or NOT rejecting the null hypothesis H_0. If p-value < 0.05then we reject H_0. If p-value > 0.05 then we do NOT reject H_0.
The variable *Net Price* (numerical variable) and *Public/Private* (categorical variable) are used in this method.

Propensity score matching

The propensity score is the probability that a unit with a particular characteristic is assigned to the treatment group as opposed to the control group. By balancing covariates between the treatment and control groups, these scores can be used to reduce or eliminate selection bias in observational studies (Stephanie, 2021).
The college type(Public/Private) will be the treatment group. Student population, alumni salary and acceptance rate will be the control groups. The net price of the college is the outcome of interest. Propensity score matching is for the college type propensity.
I will first estimate the propensity score using a logistic regression model were treatment (college types) is outcome based on the vector of covariates(student population, alumni salary and acceptance rate). The logistic regression model I will be using is:

$$log\Big(\frac{p}{1-p}\Big) = \beta_0 + \beta_1 X_{\text{Student Population}} + \beta_2 X_{\text{Alumni Salary}} + \beta_3 X_{\text{Acceptance Rate}}$$

All the predictors are numerical variables, as we have described in the data section.
Where:
p represents the probability that the college is public, as we converted public colleges to 1, in the data cleaning section.
β_0 represents the intercept of the model and is the log of odds of the college is public when *Student Population*, *Alumni Salary*, and *Acceptance Rate* equal to 0. However, in this case, the intercept makes nonsense practically as the student population cannot be equal to 0.
β_1 represents the slope of *Student Population* in the model. We expect a β_1 increase log odds of the college

is public when holding other predictors constant for every one unit increase in student population.
β_2 represents the slope of *Alumni Salary* in the model. We expect a β_2 increase log odds of the college is public when holding other predictors constant for every one unit increase in alumni salary.
β_3 represents the slope of *Acceptance Rate* in the model. We expect a β_3 increase log odds of the college is public when holding other predictors constant for every one unit increase in acceptance rate.
Now, we have the estimated score that is the probability of each observation. And then matching the observations by the nearest neighbour matching. After that, I will evaluate matching quality by comparing the means by t-test. Eventually, I will evaluate the outcomes by comparing means, running a regression on matched samples controlling for unbalanced covariates (Caetano, 2021).
Here is the mathematical model for the multiple linear regression:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,A} + \epsilon_i$$

Where:
$i = 1, \ldots, n$ where n is the number of colleges in the sample.
Y_i is the net price for the i^{th} college.
$X_{i,1}$ is the student population for the i^{th} college.
$X_{i,2}$ is the alumni salary for the i^{th} college.
$X_{i,3}$ is the acceptance rate for the i^{th} college.
$X_{i,A}$ is the college type for the i^{th} college. It equals 1 if the college type for i^{th} college is public. It equals 0 if the i^{th} college is private.
β_0 is the intercept of the private college. It represents the expected value of the net price for private college when other predictors are 0.
$\beta_0 + \beta_4$ is the intercept of public college. It represents the expected value of the net price for public college when other predictors are 0.
β_1 represents by holding other predictors constant, the number of student population increase by one unit, the expected value of the net price of colleges changes by β_1.
β_2 represents by holding other predictors constant, the alumni salary increase by one unit, the expected value of the colleges' net price changes by β_2.
β_3 represents by holding other predictors constant, the acceptance rate increase by one unit, the expected value of the colleges' net price changes by β_3.
β_4 represents by holding other predictors constant, when the college is public, the expected value of colleges' net price changes by β_4.
The mathematical model implicitly assume that a linear relationship exists in the population. Assume ϵ_i are independent and identically distributed error terms. That is $\epsilon_i \overset{iid}{\sim} Normal(0, \sigma^2)$, which implies Y_i follows Normal distribution. Another assumption is that the error are uncorrelated, namely $Cov(\epsilon_i, \epsilon_j) = 0$ or $Cov(y_i, y_j) = 0$. Last assumption is that the errors ϵ_i have a common variance σ^2. That is constant error variance, namely, homoskedasticity. (Daignault, 2021)

Results

The net price varies between public and private colleges. However, it may result from the confounding variables within the dataset. Therefore, I perform the propensity score matching method to reduce or eliminate selection bias in observational studies. After all, we will know that there will be a causal inference between college type and net cost.

Two sample t-test

I have stated in the method section that:
$H_0 : \mu_{public} = \mu_{private}$
$H_A : \mu_{public} \neq \mu_{private}$

Where μ_{public} is the population, mean of the average net price for public college, and $\mu_{private}$ is the population mean of the average net price for private college.
The difference between the sample means of the net price for public and private college is -$11653.76. The corresponding p-value is 0 indicates that we will reject the null hypothesis H_0. Thus, we have strong evidence that the population mean of the net price for public and private colleges is different.

Propensity score matching

Estimate the propensity score

Table 5: estimated coefficients of logistic regression model

term	estimate	standard error	statistic	p-value
Intercept	-5.51	1.302	-4.234	2.30e-05
student population	1.728e-04	1.425e-05	12.121	$<$ 2e-16
alumni salary	4.317e-06	1.007e-05	0.429	0.668
acceptance rate	0.041	0.007	5.896	3.73e-09

Table 5 is a summary table for the coefficient of the logistic regression model. From table 5, eventually, the estimated logistic regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.51 + 1.728\text{e-}04x_{\text{student population}} + 4.317\text{e-}06x_{\text{alumni salary}} + 0.041x_{\text{acceptance rate}}$$

where $\hat{p}$ represents the estimated probability of the college that is public.
For $\hat{\beta}_1 = 1.728e-04$, for every one-unit increase in student population, the average log odds of the college is public will increase by $1.728e-04$ when holding other predictors constant.
For $\hat{\beta}_2 = 4.317e-06$, for every one-unit increase in alumni salary, the average log odds of the college is public will increase by $4.317e-06$ when holding other predictors constant.
For $\hat{\beta}_3 = 0.041$, for every one-unit increase in acceptance rate, the average log odds of the public college will increase by 0.041 when holding other predictors constant.
As $\hat{\beta}_0 = -5.51$, that is the intercept of the model. It means when *Student Population*, *Alumni Salary*, and *Acceptance Rate* equal 0, the expected log odds of the public college is -5.51. However, for example, we will not have a student population is 0. Thus, the intercept of the model seems does not seem to make sense in real life.
Using this logistic regression model where college type is outcome based on *Student Population*, *Alumni Salary*, and *Acceptance Rate*, we estimate the propensity score now.

Evaluate quality of matching

The null hypothesis is $H_0 : \mu_{public} = \mu_{private}$
The alternative hypothesis is $H_A : \mu_{public} \neq \mu_{private}$
Where μ_{public} is the population mean of average acceptance rate for public college, and $\mu_{private}$ is the population mean of average acceptance rate for private college.
The difference between the sample means of the acceptance rate for public and private colleges is 1.83%, the corresponding p-value is 0.296 indicates that we will not reject the null hypothesis H_0. Thus, we have strong evidence that the population mean acceptance rate for public and private colleges is the same. It means that our control groups have very similar attributes after matching. Thus, the matching is reasonably successful. After matching, we have 496 matched observations by removing the unpaired observations.

Evaluate the outcomes

Table 6: estimated coefficients of multiple linear regression model

term	estimate	standard error	statistic	p-value
Intercept	8898.79	2866.40	3.10	0.00202
student population	0.05	0.02	2.24	0.02526
alumni salary	0.14	0.02	6.43	3.04e-10
acceptance rate	68.14	15.19	4.49	9.08e-06
Public/Private	-13084.89	591.37	-22.13	$<$ 2e-16

Table 6 is a summary table for the estimated coefficient of the multiple linear regression model. From table 6, eventually, the estimated multiple linear regression model is:

$$\hat{Y_i} = 8898.79 + 0.05X_{i,1} + 0.14X_{i,2} + 68.14X_{i,3} - 13084.89X_{i,A}$$

where $i = 1, ..., n$ where n is the number of colleges in the sample, $\hat{Y_i}$ is the expected net price for the i^{th} college, $X_{i,1}$ is the student population for the i^{th} college, $X_{i,2}$ is the alumni salary for the i^{th} college. $X_{i,3}$ is the acceptance rate for the i^{th} college. $X_{i,A}$ is the college type for the i^{th} college. It equals 1 if it's the public college. It equals 0 if it's the private college.
By holding other predictors constant, the student population increase by one unit, the expected value of the net price of colleges increase by around \$0.05.
By holding other predictors constant, the alumni salary increase by one unit, the expected value of the net price of colleges increases by around \$0.14.
By holding other predictors constant, the acceptance rate increases by one unit, and the expected value of colleges' net price increases by around \$68.14.
By holding other predictors constant, when the college is public, the expected value of colleges' net price decrease by around \$13084.89.
As $\hat{\beta}_0 = 8898.79$, It means the expected value of the net price for private colleges is 8898.79 when other predictors are 0. It does not seem to make sense in real life as the student population, alumni salary, and acceptance will not be 0 practically.
For $\hat{\beta}_0 + \hat{\beta}_4 = -4186.11$. It represents the expected value of the net price for public college is \$-4186.11, when other predictors are 0. It does not seem to make sense in real life as the net price will not be negative practically.
Altogether, we know that the estimated model when the college is **private** would be:

$$\hat{Y_i} = 8898.79 + 0.05X_{i,1} + 0.14X_{i,2} + 68.14X_{i,3}$$

where $\hat{Y_i}$ is the expected net price of colleges, $X_{i,1}$ is the student population of colleges, $X_{i,2}$ is the alumni salary of colleges. $X_{i,3}$ is the acceptance rate of colleges.
Then, we know that the estimated model when the college is **public** would be:

$$\hat{Y_i} = -4186.11 + 0.05X_{i,1} + 0.14X_{i,2} + 68.14X_{i,3}$$

where $\hat{Y_i}$ is the expected net price of colleges, $X_{i,1}$ is the student population of colleges, $X_{i,2}$ is the alumni salary of colleges. $X_{i,3}$ is the acceptance rate of colleges.
Thus we know that by considering the similar condition(control group), holding these conditions constant, the public colleges have less \$13084.9 net price than the private colleges.
The p-value for $H_0 : \beta_4 = 0$, $H_A : \beta_4 \neq 0$ is close to 0. The common significant level is 0.05. It means that when the p-value is lower than 0.05, we have strong evidence against that $H_0 : \beta_4 = 0$. We have strong evidence to support that the college type would change the impact of other predictors on colleges' net price. As the control groups have similar attributes after propensity score matching, only the college types vary. We have strong evidence to show that the college types have causal inference with the colleges' net price. Thus, since the college is private, it has a higher net price than the public college. More specifically, on average higher than \$13084.9.

Conclusions

As specified in the introduction, the research question is: What is the relationship between the college type and its net costs? I hypothesize that even though subtracting the financial aid, on average, private colleges still have higher net costs than public colleges as private colleges rely on student tuition fees, alumni, and endowments to fund their academic programs. From the results section, I find out that the results indeed captures the hypotheses. The results suggest a causal inference with the net price of the colleges. Since the college is private, it has a higher net price than the public college on average. More specifically, on average higher than $13084.9.

Taken together, I used the practical rationale to find out the estimated multiple linear regression model for this report. That is $\widehat{Y_i} = 8898.79 + 0.05X_{i,1} + 0.14X_{i,2} + 68.14X_{i,3} - 13084.89X_{i,A}$, where $i = 1, \ldots, n$ where n is the number of colleges in the sample, $\widehat{Y_i}$ is the expected net price for the i^{th} college, $X_{i,1}$ is the student population for the i^{th} college, $X_{i,2}$ is the alumni salary for the i^{th} college. $X_{i,3}$ is the acceptance rate for the i^{th} college. $X_{i,A}$ is the college type for the i^{th} college.

Practically, when we find out that by holding other predictors constant, the number of student population increase by one unit, the expected value of the net price of colleges increase by around $0.05. By holding other predictors constant, the alumni salary increases by one unit, and the expected value of colleges' net price increases by around $0.14. By holding other predictors constant, the acceptance rate increase by one unit, the expected value of the net price of colleges increases by around $68.14. By holding other predictors constant, when the college is public, the expected value of colleges' net price decrease by around $13084.89. Thus we know that by considering the similar condition(control group), holding these conditions constant, on average, the public colleges has less $13084.9 net price than the private colleges.

Besides, I have a detailed data collection and cleaning process with variable descriptions in the data section. Also, I want to emphasize several critical results in the data section. From figure 1, we can see that we expect to approximately cost $20,000 per year in most of the colleges. Besides, we found that the net price of college is symmetric. From figure 2, we found that private colleges are about 1.5 times that of public colleges in America. From figure3, we would expect a higher net price for private colleges and fluctuate more significantly than the public colleges.

Overall, this report uses propensity score matching to find out the causal inference: since the college is private, on average, it has a higher net price, by $13084.9, than the public college. It suggests that those who want to apply for a college if a severe financial crisis could choose a public college or a private school with financial aid.

Weaknesses

The data is collected from 2019(pre-COVID-19 period). The net cost may differ after the COVID-19 pandemic. Moreover, the dataset only contains the top 650 colleges in America. Thus the data may not be a perfect representative sample.

Besides, propensity score matching(PSM) could increase imbalance, inefficiency, model dependence, and bias. The weakness of PSM comes from its attempts to approximate an utterly randomized experiment rather than, as with other matching methods, a more efficient, entirely blocked randomized experiment (King, 2019).

Next Steps

It could be interesting to find other confounding variables other than student population, alumni salary and acceptance rate since other variables may also impact the net price of different colleges. Alternatively, using another method other than propensity score matching to investigate the research question as PSM has some weaknesses stated in the weakness section. However, I cannot discuss it here since I do not have the data regarding the other confounding variables.

Discussion

Throughout the report, I first hypothesize that even though subtracting the financial aid, on average, the private colleges still have a higher net cost than public colleges in the introduction. Using the method such as propensity score matching found out that there is a causal inference with the colleges' net price. Since the college is private, on average, it has a higher net price than the public college. More specifically, on average higher than $13084.9. Moreover, I include a data section that includes some numerical and graphical analyses of net price and visualization of the proportion of private and public colleges.

Bibliography

All analysis for this report was programmed using `R version 4.1.2`.
1. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: December 17, 2021)
2. *Assumptions of logistic regression.* Statistics Solutions. (2021, August 11). Retrieved December 6, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/.
3. Beers, B. (2021, December 7). *What P-value tells us*. Investopedia. Retrieved December 12, 2021, from https://www.investopedia.com/terms/p/p-value.asp.
4. Caetano, S. (2021). *Week9-Propensity Score Matching.pdf*. Retrieved December 12, 2021, from https://q.utoronto.ca/courses/236142/pages/w9-slides-and-videos?module_item_id=3135427
5. Cammilleri, C. (2020, August 07). *America's top college Rankings 2019 (Forbes)*. Retrieved December 3, 2021, from https://www.kaggle.com/chris95cam/forbes-americas-top-colleges-2019
6. Conklin, J., Coudriet, C., & Howard, C. (n.d.). *America's Top Colleges 2019*. Retrieved December 3, 2021, from https://www.forbes.com/top-colleges/#2a7e02771987
7. Daignault, K. (2021, October). *Module 3 Asynchronous Material: Assumptions and Properties of Estimators*. Lecture. Retrieved December 12, 2021, from https://q.utoronto.ca/courses/236123/pages/module-3-asynchronous-material?module_item_id=3011149
8. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.
9. Gergely Daróczi and Roman Tsegelskyi (2021). *pander: An R 'Pandoc' Writer*. R package version 0.6.4. https://CRAN.R-project.org/package=pander
10. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: December 17, 2021)
11. King, G. & Nielsen, R, 2019, *'Why Propensity Scores Should Not Be Used for Matching'*,Political Analysis.
12. Kreznar, C. (2021, September 8). *How we Rank America's top colleges.* Forbes. Retrieved December 3, 2021, from https://www.forbes.com/sites/christiankreznar/2021/09/08/how-we-rank-americas-top-colleges/?sh=3550142c43e0.
13. Powell, F., Kerr, E., &; Wood, S. (2021, September 17). *What you need to know about college tuition costs*. Retrieved December 14, 2021, from https://www.usnews.com/education/best-colleges/paying-for-college/articles/what-you-need-to-know-about-college-tuition-costs
14. R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
15. Stephanie. (2021, April 2). *Propensity score matching: Definition & Overview*. Statistics How To. Retrieved December 13, 2021, from https://www.statisticshowto.com/propensity-score-matching/.
16. Wickham et al., (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix

A1: Ethics Statement

America's top colleges' dataset used in this report is free and under public domain license (Cammilleri, 2020). I am honest in reporting results, and there is no p-hacking throughout the report, never deliberately trying to make the p-value significant by human intervention such as changing or deleting the data.
I respect and acknowledge the contributions and intellectual property of others. I have appropriately cited them in the bibliography and used inline APA format citations.

A2: Materials

Here is a glimpse of the original dataset:

```
## Rows: 650
## Columns: 17
## $ Rank                    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14~
## $ Name                    <chr> "Harvard University", "Stanford University", ~
## $ City                    <chr> "Cambridge", "Stanford", "New Haven", "Cambri~
## $ State                   <chr> "MA", "CA", "CT", "MA", "NJ", "PA", "RI", "CA~
## $ Public.Private          <chr> "Private", "Private", "Private", "Private", "~
## $ Undergraduate.Population <dbl> 13844, 8402, 6483, 4680, 5659, 13437, 7390, 1~
## $ Student.Population      <dbl> 31120, 17534, 12974, 11466, 8273, 25367, 1009~
## $ Net.Price               <dbl> 14327, 13261, 18627, 20771, 9327, 24242, 3020~
## $ Average.Grant.Aid       <dbl> 49870, 50134, 50897, 43248, 48088, 44801, 424~
## $ Total.Annual.Cost       <dbl> 69600, 69109, 71290, 67430, 66150, 71715, 710~
## $ Alumni.Salary           <dbl> 146800, 145200, 138300, 155200, 139400, 13390~
## $ Acceptance.Rate         <dbl> 5, 5, 7, 7, 6, 9, 8, 8, 10, 10, 13, 8, 17, 7,~
## $ SAT.Lower               <dbl> 1460, 1390, 1460, 1490, 1430, 1420, 1405, 153~
## $ SAT.Upper               <dbl> 1590, 1540, 1580, 1570, 1570, 1560, 1570, 159~
## $ ACT.Lower               <dbl> 32, 32, 32, 33, 31, 32, 31, 34, 31, 30, 31, 3~
## $ ACT.Upper               <dbl> 35, 35, 35, 35, 35, 35, 35, 35, 35, 34, 34, 3~
## $ Website                 <chr> "www.harvard.edu", "www.stanford.edu", "www.y~
```

Here is a glimpse of the cleaned dataset:

```
## Rows: 632
## Columns: 7
## $ Rank                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ Name                 <chr> "Harvard University", "Stanford University", "Yal~
## $ `Public/Private`     <chr> "Private", "Private", "Private", "Private", "Priv~
## $ `Student Population` <dbl> 31120, 17534, 12974, 11466, 8273, 25367, 10095, 2~
## $ `Net Price`          <dbl> 14327, 13261, 18627, 20771, 9327, 24242, 30205, 2~
## $ `Alumni Salary`      <dbl> 146800, 145200, 138300, 155200, 139400, 133900, 1~
## $ `Acceptance Rate`    <dbl> 5, 5, 7, 7, 6, 9, 8, 8, 10, 10, 13, 8, 17, 7, 16,~
```

Supplementary Plots and Tables

Table 7: matched colleges with similar attributes

Rank	Name	Public/Private	Student Population	Net Price	Alumni Salary	Acceptance Rate	.fitted	cnts
25	Swarthmore College	0	1630	20511	123200	11	0.01416	1
53	United States Coast Guard Academy	1	1044	0	118100	15	0.01476	1
66	United States Merchant Marine Academy	1	975	6758	140700	22	0.02133	1
52	Carleton College	0	2155	27898	109900	21	0.02195	1
24	United States Naval Academy	1	4526	0	152800	8	0.02324	1
406	New Mexico Institute of Mining and Technology	1	2009	13694	120400	22	0.02331	1