BUAN 6392
Causal Analytics and A/B Testing Project Report

Group 5: Yu-Chien Ho, Xuan Fan, Xiaojia Zhang

Abstract

Due to the Pandemic situation, most Universities transferred their classes to online mode for safety purposes. Which gave the students freer time since the remote teaching style are more flexible than traditional teaching style, some students took the advantage by taking extracurricular classes online to increase their competitiveness. Standing in a stronger position when entering the job market.

Our research focused on whether taking extracurricular classes outside school affect students' GPA when compare to the students who only taking classes within the University. To accomplish the research, we applied the UTD IRB board for approval to collecting data on individual subject, analysis the survey results to support the conclusion of the project.

To retrieve the results, we used power analysis method for conducted sufficiency, propensity score matching (PSM) to match similar individuals and Welch's t-test to test whether two populations have equal means.

By researching whether taking extracurricular classes impact student's GPA, it becomes evident that it has influence on their GPA based on the data and experience results we had, taking extra online classes action reflecting their attitude towards academic achievement, and the results are obvious which are demonstrate on the GPA compassion.

Introduction

The world is experiencing a once-in-a-lifetime pandemic, causing untold human suffering and death, unraveling of social relationships, and robbing individuals of livelihoods and countries of prosperity. The coronavirus pandemic has strained health systems, revealed unconscionable inequalities, and upended international institutions.

The University of Texas at Dallas is dedicated to operating safely during the COVID-19 pandemic by providing comprehensive guidance and essential resources and transfer most classes online to minimize the spread of COVID-19 and protect the safety of all students and faculty members.

The online classes have a more flexible time schedule, which give students more opportunities to arrange their time. Students are not able to go to school or use any resources from the library which indicate that online resources became relatively important. More and more online academy came up with some online learning courses, some students chose to take extracurricular classes outside school like Udemy, to gain more skills and learn knowledge for improvements, enhance their ability to their relatively major without massive tuition fee.

Therefore, we would like to use this experiment to find out whether students taking online courses will potentially help them increasing their GPA.

The research was focusing specifically on the extra online class's influence on student grades, will the student grades increase or decrease when compare with those students who didn't take any classes outside school, focused mainly on schoolwork.

We applied to the UTD IRB board and applied for their approval since we are collecting data on individual subjects. We all got the human subjects training certificates first and then described all the variables that we would be collecting in our survey in the IRB application. Also, we submitted recruitment scripts and consent form as our supporting documents. After we got the approval from the IRB board, we sent out our survey to students.

After collected the data from the survey we applied the power analysis method, propensity score matching and Welch's t-test to get the result. The t-test confirms that for the full sample, students who take extra online courses have a greater likelihood of receiving higher grades on average than their school course only counterparts.

Question

Whether taking extracurricular online courses has effect on GPA?

Experiment Design

In our research, as we mentioned in the previous part, we are curious about whether extracurricular classes outside school affect students' GPA when compared to the students who only take classes within the University. Therefore, we would like to conduct an A/B testing which set our control group as students who had never taken extracurricular classes outside of school and the treatment group as students who actively take extracurricular classes.

As we know, conducting A/B testing must have control and treatment groups to be randomly assigned. However, due to lack of resources to set up a tracking system in particular online courses websites, the only way to let us conduct this research is to use Quasi Experiment which we will send out an online survey through E-learning that we created on Qualtrics Survey Solution to the students with different majors who enrolled in University of Texas at Dallas.

Before we start our A/B testing analysis, we firstly want to know the data that we have collected was significantly enough to conduct t-test. Therefore, we will use power analysis to check our observations' size and we expected that the size of each group will be equal, also, the total size of this analysis will be smaller than 100 because we only have 100 observations after the collecting process.

After making sure the sample size is enough, we will then conduct the Propensity Score Matching to match similar individuals in each group and we expect that the matched sample size will be greater than the effect sample size which can lead us to continue conducting our hypothesis testing.

Last but not least, we want to formulate the null and alternative hypothesis. In our case, we are asking students who have taken extracurricular online courses outside of school and expecting that this difference will appear in their GPA. Thus, the null hypothesis, in this case, should be formulated as follows:

- H_0- Null Hypothesis: taking extracurricular online courses doesn't have an effect on GPA.
- Alternative Hypothesis: taking extracurricular online courses has an effect on GPA.

The A/B testing will aim to reject the null hypothesis with a significant level of p-value to be set as 0.05.

Data Collection

Since collecting information will be sensitive about students' performance, which means we need to apply to UTD IRB board and get the approval of collecting individual subjects. The process would be navigating to the Office of Research Integrity and Outreach's website and submit our research proposal, the certificates of human training, and finally the recruitment script of how we would send out our survey.

Once we get the approval from IRB, we then start to send out our survey through e-learning to our classmates who have enrolled in Causal Analytics and A/B Testing. In order to obtain more students to finish our survey, we also send out our survey through WeChat and line groups of which the members are all the students from UTD.

Our survey includes several questions regarding this individual's behavior of using extracurricular online courses, for instance, hours of using this resource. We also include some basic questions about what's major, degree, GPA and enroll semester of this individual.

Data Description

Major: char, unit's major, ex: Business Analytics

Degree: char, unit's degree, ex: Master

Enroll_sem: char, the semester and year when unit enrolled in, ex: Fall 2019

GPA: float with two decimals, the overall current GPA after Summer 2020, ex: 3.67

Gender: factor, binary variable, 1 for female, 0 for male

Age: factor, integer interval, age range, ex: 20 ~ 25

Resources: factor, binary variable, if the unit uses learning resource, 1 for yes, 0 for no

Hours: factor, time interval (hour range), how many hours does the unit take, ex: <1 hour

Job_intern: factor, binary variable, whether the unit has job/intern or not, 1 for yes, 0 for no

Power Analysis

Before we conduct the hypothesis testing, we should figure out if our observations are sufficiently enough to conduct the experiment. Hense, we run pwr.t.test in r:

```
     Two-sample t test power calculation

              n = 38.9
              d = 0.745
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

As we can see from the output above, the effect sample size is 38.9 for each group (total 77.8). Since we obtained almost 100 observations, our experiment can be conducted sufficiently.

Propensity Score Matching (PSM)

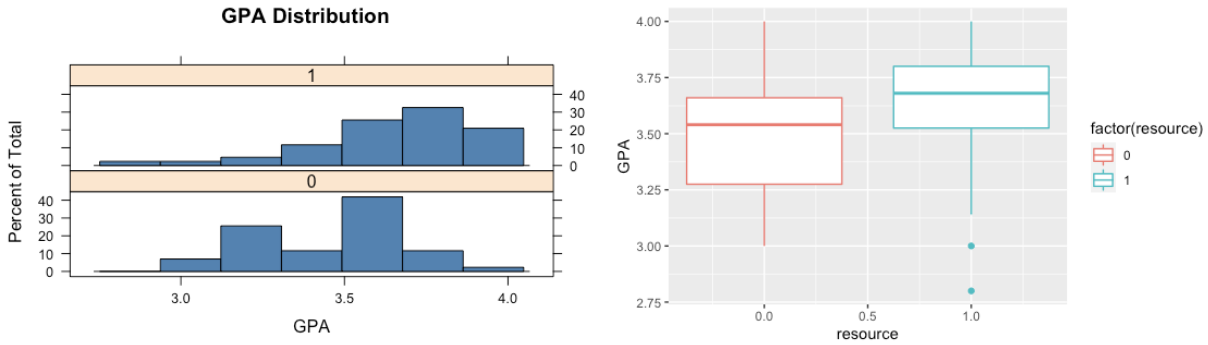
Since the treatment group and control group are not randomly assigned, we need to utilize Propensity Score Matching (PSM) approach to match similar individuals in each group. Therefore, we run PSM using MatchIt in r:

```
Sample sizes:
          Control Treated
All            43      60
Matched        43      43
Unmatched       0      17
Discarded       0       0
```

According to the output, there are 43 individuals in total are matched. Additionally, 43 is larger than 18.7, the effect sample size. Therefore, our experiment can be conducted sufficiently.

Welch's t-test

Before we conduct hypothesis test, let's look at the histogram and boxplot of each group:



For the t-test to be valid, the data in each group should be approximately normal. If the distributions are different, minimally Welch's t-test should be used.

- H_o - Null Hypothesis: taking extracurricular online courses hasn't an effect on GPA.
- H_a - Alternative Hypothesis: taking extracurricular online courses has an effect on GPA.

```
        Welch Two Sample t-test

data:  GPA by resource
t = -5, df = 83, p-value = 0.0000005
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.412 -0.192
sample estimates:
mean in group 0 mean in group 1
           3.40            3.71
```

From the result of Welch's t-test, the p-value less than 5% (assume the significant level is 5%). Hence, we are confident to reject the null hypothesis and draw the conclusion that taking extracurricular online courses has an effect (improvement) on GPA.

Conclusion

After preparing for three weeks, our experiment was conducted successfully as scheduled and the result is the same as what we expected. To some extent, taking extracurricular online courses do improve students' GPA.

There are some outliers in the dataset. Some of them didn't take any extracurricular online courses, however, they still have relatively high GPA and work simultaneously. These situations occur due to individual differences. They may have several internship experiences before so that it's easier for them to get full-time jobs or more internships. Additionally, these hands-on experiences improve students' abilities to complete courses with high quality.

Since we're not able to randomly assign students to utilize or not utilize online courses, the approach we used is Propensity Score Matching (PSM). Treatment group is composed of students who are utilizing online resources. Control group is composed of students who are not utilizing online resources. In order to find the effect size for our experiment, we performed Power Analysis and found out that the effect size should be 38.9 or higher for each group. Since the matched sample size is 43, our experiment can be carried out sufficiently. For t-test to be valid, the distributions of both treatment group and control group are supposed to be approximately normal, otherwise we should perform Welch's t-test (under 95% confidence interval). As we can see in the histograms, the two distributions are different and the one without resources are not normally distributed. Therefore, we conducted Welch's t-test, and ultimately, we get p-value is less than the significant level.

Our experiment shows that taking extracurricular online courses improve students' GPA. As a matter of the fact, not only students but also industrial professionals utilize online learning websites for self-improvement, especially during this pandemic.

Appendix

R Code

```
library(pacman)

pacman::p_load(readxl, ggplot2, dplyr, caret, pwr, MatchIt, lattice)

options(scipen = 999)

options (digits = 3)

options(warn=0)

set.seed(123)

#----------------------Load data ----------------------------

mydata <- as.data.frame(read_excel("Project Survey.xls"))

View(mydata)

#-------------------Data Preparation -------------------------

str(mydata)

#Average GPA for each major

mydata %>%

  group_by(major) %>%

  summarise_at(vars (GPA), funs (mean (., na.rm = TRUE)))

#Convert gender, resource, job/intern into binary variables. Label encoding.

mydata$gender <- ifelse(mydata$gender == "Female", 1, 0)

table(mydata$gender)

mydata$resource <- ifelse(mydata$resource == "Yes", 1, 0)

table(mydata$resource)

mydata$job_intern <- ifelse(mydata$job_intern == "Yes", 1, 0)

table(mydata$job_intern)

#Convert major, degree, enroll_sem

table(mydata$major)

mydata$major %>%

  factor (levels = c ("BA", "CS", "ECN", "FIN", "ITM", "SCM"))
```

```
table(mydata$degree)

mydata$degree %>%

  factor (levels = c ("Bachelor", "Master", "PhD"), labels = c ('B', 'M', 'P'))

table(mydata$enroll_sem)

mydata$enroll_sem %>%

  factor (levels = c ("Fall 2017", "Fall 2018", "Fall 2019", "Spring 2018", "Spring 2019", "Spring 2020"), labels = c (1, 2, 3, 4, 5, 6))

table(mydata$age)

mydata$age %>%

  factor (levels = c ("19 or less", "20 ~ 25", "25 ~ 30", "30 +"), labels = c (0, 1, 2, 3))

table(mydata$hours)

mydata$hours %>%

  factor (levels = c("0", "< 1hr", "1~2 hrs", "2~3 hrs", "4 hrs +"),

        labels = c (0, 1, 2, 3, 4))

#----------------------Set Treatment Group & Control Group ------------------------

#Treatment Group: individuals who use online resources, resource = 1

treat <- mydata %>% subset(resource == 1)

M1 = mean(treat$GPA) #3.65

S1 = sd(treat$GPA) #0.266

#Control Group: individuals who don't use online resources, resource = 0

ctrol <- mydata %>% subset(resource == 0)

M2 = mean(ctrol$GPA) #3.47

S2 = sd(ctrol$GPA) #0.229

#-------------------------------------Power Analysis----------------------------------------

cal_d = (mean(treat$GPA) - mean(ctrol$GPA))/sqrt(((sd(treat$GPA)^2) + (sd(ctrol$GPA)^2))/2)

pwr.t.test(n = NULL, d = cal_d, sig.level = 0.05, power = 0.9,

          type = "two.sample", alternative = "two.sided")
```

```
#The power is the likelihood of finding statistical significance.
#--------------------------------Propensity Score Matching--------------------------------
mymatch = matchit(resource ~ major + degree + enroll_sem + gender + age + GPA + hours +
                    job_intern, data = mydata, method = "nearest", ratio = 1)
summary(mymatch)
#-----------------------------------------------Analysis-----------------------------------------------
#Obtain the matched data
matched_data <- match.data(mymatch)
View(matched_data)
 #Histogram and Boxplot for each group
histogram(~ GPA | factor(resource), data = matched_data, col = "steelblue",
         main = "GPA Distribution")
matched_data %>% ggplot(aes(x = resource, y = GPA, group = factor(resource))) +
                geom_boxplot(aes(colour = factor(resource)))
#Welch's t-test
t.test(GPA ~ resource, data = matched, var.equal = FALSE, conf.level = 0.95)
```

Dataset

BUAN6392 Project Survey_group5.csv