# Yao_Candice_final

Candice Yao(jy3440)

Due December 20, 2023 at 5pm

## Instructions

*You should submit your write-up (as a knitted .pdf along with the accompanying .rmd file) to the course website before 5pm EST on Wednesday, Dec 20th Please upload your solutions as a .pdf file saved as* `Yourlastname_Yourfirstname_final.pdf`*.In addition, an electronic copy of your .Rmd file (saved as* `Yourlastname_Yourfirstname_final.Rmd`*) should accompany this submission.*

*Late finals will not be accepted,* ***so start early and plan to finish early****.*

*Remember that exams often take longer to finish than you might expect.*

*This exam has* ***3*** *parts and is worth a total of* ***100 points****. Show your work in order to receive partial credit.*

*Also, we will penalize uncompiled .rmd files and missing pdf or rmd files by 5 points.*

*In general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions we have asked, you will not receive points when you demonstrate knowledge about questions we have not asked, and you will lose points when you make inaccurate statements (whether or not they relate to the question asked). Be careful, however, that you provide an answer to all parts of each question.*

*You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself. You are prohibited from corresponding with any human being regarding the exam (unless following the procedures below).*

*The TAs and I will answer clarifying questions during the exam. We will not answer statistical or computational questions until after the exam is over. If you have a question, send email to all of us. If your question is a clarifying one, we will reply. Do not attempt to ask questions related to the exam on the discussion board.*

# Problem 1 (100 points)

In this problem, you will examine whether family income affects an individual's likelihood to enroll in college by analyzing a survey of approximately 4739 high school seniors that was conducted in 1980 with a follow-up survey taken in 1986.

This dataset is based on a dataset from

> Rouse, Cecilia Elena. "Democratization or diversion? The effect of community colleges on educational attainment." Journal of Business & Economic Statistics 13, no. 2 (1995): 217-224.

The dataset is `college.csv` and it contains the following variables:

- `college` Indicator for whether an individual attended college. (Outcome)
- `income` Is the family income above USD 25,000 per year (Treatment)
- `distance` distance from 4-year college (in 10s of miles).
- `score` These are achievement tests given to high school seniors in the sample in 1980.
- `fcollege` Is the father a college graduate?
- `tuition` Average state 4-year college tuition (in 1000 USD).
- `wage` State hourly wage in manufacturing in 1980.
- `urban` Does the family live in an urban area?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dagitty)
library(broom)
library(dplyr)
```

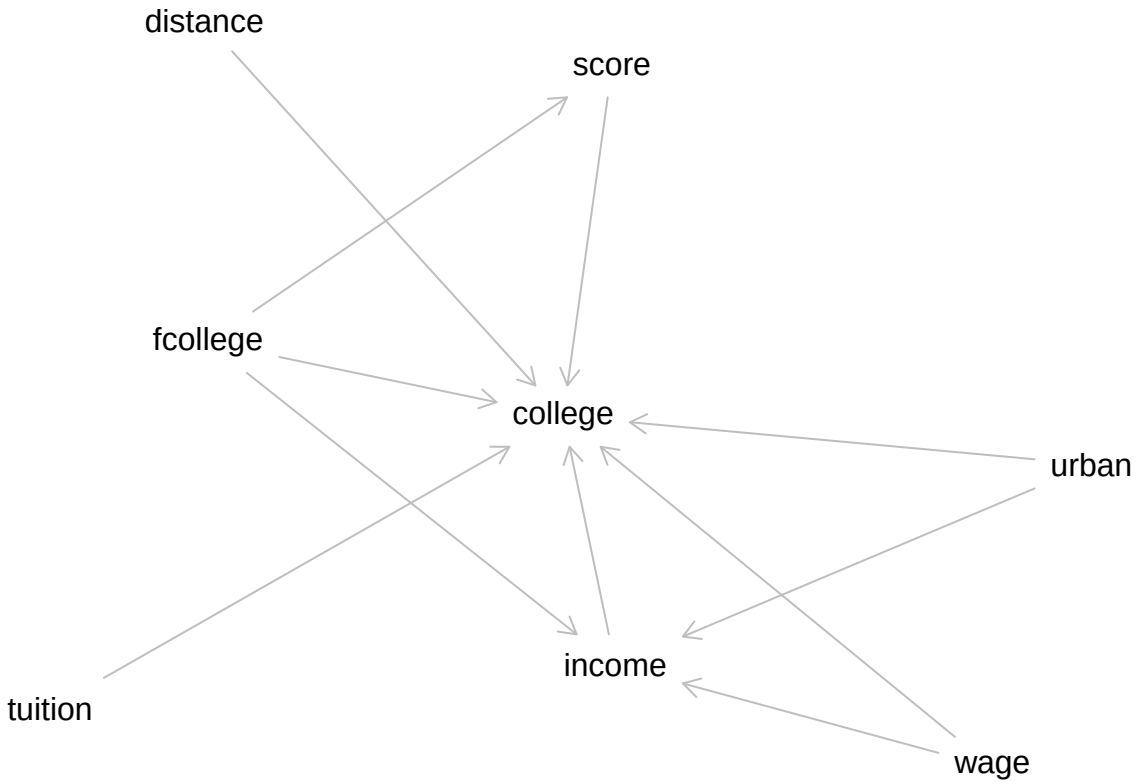
## Question A (35 points)

Draw a DAG of the variables included in the dataset, and explain why you think arrows between variables are present or absent. You can use any tool you want to create an image of your DAG, but make sure you embed it on your compiled .pdf file. Assuming that there are no unobserved confounders, what variables should you condition on in order to estimate the effect of the treatment on the outcome, according to the DAG you drew? Explain your decision in detail. In your explanation, provide a definition of confounding.

```
rm(list = ls())
college <- read.csv("college.csv", na.strings = c(""))
```

```
g <- dagitty("dag{
income -> college;
distance -> college;
fcollege -> college;
urban -> income;
urban -> college;
score -> college;
tuition -> college;
wage -> income;
wage -> college;
fcollege -> score;
fcollege -> income
}")

plot(g)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```




In the context of this data, there is an arrow between income(treatment) and college(outcome), since the purpose is to estimate how has income affected educational outcome of individuals.

In addition to this, I identify three confounders, i.e. covariates that affect both selection into treatment and outcomes: fcollege, wage, and urban.

- Having a college degree is often associated with higher likelihood of white-collar employment, which generally receives more payment than blue-collar jobs, thereby I posit that having a father who has a

college degree will likely increase a student's family income. In addition, a father who has a college degree might have higher educational expectation of their child, be able to devote more resources to this end. Therefore fcollege is linked to both treatment and outcome.

- State manufacturing wage is another possible confounding factor. First, it might directly affect some people's familial income if they have family members working in manufacturing. Alternatively, if we take it as a proxy for the overall state-level economic situation, it can be affecting units' overall familial incomes despite their family members' careers. Its influence over the outcome can also be two-folded. First, it could increase the opportunity cost of students to attend college. In other words, if the manufacturing wage is exceptionally high, individuals might be more likely to join the workforce without attending college. Similarly, if this factor is indicative of the overall economic situation, higher manufacturing wage may represent a flourishing economy and an optimistic job market, driving people to pursue higher education and white-collar jobs.
- Urban(whether the family lives in an urban area) is another confounder because I assume that living in a city will expose one's parents to more career opportunities since more companies are located in the city and thus increase one's familial income. This factor also potentially increases one's chances to get into a college since better educational resources often cluster in the city. A student living in the city might also also exposed to a more divserse set of extracurricular activities, and thereby increasing the likelihood of them to get into college.

The score is an mediator between fcollege and college, since I believe that a student is more likely to receive academic support from their family if their father has a college degree. As a result, they might have a higher score at the senior year achievement test. This consequently affects one's college admissions since the test score can be taken into consideration by the admission officers. Or, this factor could be seen as a proxy for the students' overall GPA, which will affect their college admission result.

Finally, there are two variables that have arrows pointing to college: distance and tuition. Distance could affect college attendance in two ways: first, if one lives closer to a campus, they could enjoy access to more educational resources, which could be beneficial to their college admission(think of workshops, libraries, and study spaces). Next, given that one is admitted, they might be more inclined to enroll if they live closer to the college and therefore have to pay less for relocation. Tuition affects college attendance since higher tuition could drive individuals from low-income background to abandon college admission, or decline offer even if accepted. However, this affect is expected to be heterogeneous for individuals of different economic backgrounds, since individuals from high-income families might be more resilient to higher tuition.

I believe that distance and tuition do not have any arrow pointing to the treatment because tuition often hinges on universities' financial expenditures and funding in specific years, which is irrelevant to most people's familial income - even for college faculties, an increase of the tuition does not equate a rise in their salaries. Next, one's distance to a campus might not affect one's familial income, since universities are scattered at different locations, ranging from cities, suburbs, to rural areas. Therefore, I argue that while living closer to some particular university(*for instance, NYU*) might be associated with higher income, simply living closer to a university give us little to none information regarding income.

According to this DAG, assuming that there are no unobserved confounders, we need to condition on fcollege, wage, and urban to estimate the treatment effect.

## Question B (35 points)

Choose one of the methodologies we learned in class to calculate a causal effect under conditional ignorability. What estimand are you targeting and why? Explain why you made your choice, and discuss the assumptions that are needed to apply your method of choice to this dataset. State if and why you think these assumptions hold in this dataset. In addition, choose a method to compute variance estimates (i.e., robust standard errors or bootstrapping), and discuss the reasons behind your choice in the context of this dataset.

### 1. Chosen Method

```
# check covariate balances between treatment(income beyond 25k) and control(income below 25k)
balance= college %>% group_by(income) %>%
  summarise(score=mean(score),
            fcollege_yes=sum(fcollege =='yes')/n(),
            fcollege_no=sum(fcollege =='no')/n(),
            urban_yes=sum(urban=='yes')/n(),
            urban_no=sum(urban=='no')/n(),
            wage=mean(wage),
            distance=mean(distance),
            tuition=mean(tuition))

print(balance, width = Inf)
```

```
## # A tibble: 2 x 9
##   income score fcollege_yes fcollege_no urban_yes urban_no  wage distance
##   <lgl>  <dbl>        <dbl>       <dbl>     <dbl>    <dbl> <dbl>    <dbl>
## 1 FALSE   49.9        0.117       0.883     0.252    0.748  9.44     1.92
## 2 TRUE    53.3        0.434       0.566     0.185    0.815  9.65     1.51
##   tuition
##     <dbl>
## 1   0.803
## 2   0.843
```

By conducting a preliminary balance check, I detected visible covariates imbalances between the treatment and the control group. The treated units - students whose family income is beyond 25,000 usd - were much more likely to have a father who had attained college degree compared to the control group - students whose family income was below 25.000 usd. The treated group also scored higher on average(53.33) in the senior-year achievement exams than the control group(49.9), and lived at a closer distance(1.512) to a 4-year college than the control group(1.920). In addition, there were surprisingly more controlled units living in the urban area than the treated units - this strikes me as a counter-intuitive observation since living in the cities could be associated with higher living costs, and thus requires having higher family income. Besides, the average state 4-year tuition and the state wage for manufacturing is higher for those treated units.

Overall, all covariates in this data are imbalanced to different extents between the treatment and the control group. Since some covariates are likely to be associated with both the treatment and outcome(i.e. confounders), imbalances as such could introduce bias to our estimation of treatment effect of income on educational outcomes, since the difference in outcome could be due to the differences in those covariates instead of the treatment.

I chose to estimate ATE(average treatment effect) with an inverse propensity weighting(IPW) estimator. Propensity score is the score for an individuals to be selected into treatment, given their covariates. To construct this estimator, I will first calculate a propensity score for each unit and weighting the data with their propensity scores. By upweighting underrepresented units in both groups, the distribution of the covariates at different levels will move closer to what we would expect to see in a randomized experiment. Therefore, conditioning on the propensity score will be sufficient to gives us $Y_i(1), Y_i(0) \perp D_i \mid X_i = x$ for all x and d $(assignment of treatment is independent of covariates), so I can proceed to estimate the average treatment effect under conditional ignorability. The ATE will be computed by calculating difference in difference between the weighted treated group and the weighted control group.

**2. Estimand**

- ATE is the targeting estimand, because my chosen estimator is the Horvitz-Thompson Estimator, also known as the Inverse-probability-weighted(IPW) estimator. The propensity score will be calculated for all units in our sample, and it will be used to construct a weighted sample with balanced covariates across the control and treatment groups. With the balanced data, we are able to estimate a more precise ATE across the sample without conditioning on any covariates, since the variation of covariates are already accounted for by the propensity score weighting.

**3. Methods to Compute Variance: bootstrapping**

- Bootstrapping will be used to construct variances and the 95% confidence interval of my estimate. It is chosen because we are estimating the variances of a constructed estimator. Bootstrapping will be able to generate different weights from sampling with replacement for multiple times and, account for the randomness in our data, and thus be able to provide a more robust and precise estimate of the variances and the confidence intervals. In addition, bootstrapping has been used as a standard means to compute variances for propensity score involved estimators.

**4. Assumptions**

- The choice of IPW estimator for ATE is based on my assumption of conditional ignorability that there is no unobserved confounder that the model does not account for. In other words, all variables that is linked to both the treatment(income) and the outcome(education) are observed and taken into consideration by the model. With this assumption, by conditioning on the confounders through controlling for the propensity score, we break the backdoor paths, so we can sufficiently say that the treatment is approximating random and identify a treatment effect. In practice, I assume that there are only two confounders as shown in my previous DAG - fcollege and urban - and there is no other unobserved confounders, and construct the IPW estimator with this two confounders. It is noteworthy that this is a strong assumption, especially given the fact that IPW estimator is very sensitive to model misspecification(if the necessary confounders are not included, or the wrong confounders are selected) and extreme weight values(if some of the weights are biased, it could severely jeopardize the overall estimate).
- Despite the conditional ignorability assumption, we also assume conditional positivity, which means that given our set of observed covariates, the possibility of being selected into treatment is between 0 and 1 for all covariates and across different levels of each covariate. Later in question B, I test this assumption by locating the maximum and minimum of the units' propensity scores, and found that this assumption indeed holds since all values are between 0.12 to 0.70 - every individual has a certain level of possibility to be selected into treatment, despite the possibility varies.

## Question C (30 points)

Using the methodology you chose in Question B to control for the confounders you have selected in Question A, as well as the relevant R packages, provide your estimate of the causal effect of the treatment on the outcome. Using your variance estimator of choice, report standard errors and 95% confidence intervals around your estimates. Interpret your results and discuss both their statistical significance and their substantive implications. Be as specific and detailed as possible.

```
# convert all variables into numerical variable for easier analysis
college <- college %>% mutate(
    income = as.numeric(income),
    college = as.numeric(college),
```

```
    fcollege = if_else(fcollege == 'yes', 1, 0),
    urban = if_else(urban == 'yes', 1, 0))
```

```
diff_in_means_naive = mean(college$college[college$income==1])-
                      mean(college$college[college$income==0])
diff_in_means_naive
```

```
## [1] 0.2034
```

Before I construct my propensity score model, I compute a raw difference in mean - this value is what we might have gotten as the estimate if we did not use the propensity scores method to control for potential confounding between the variables with selection into treatment.

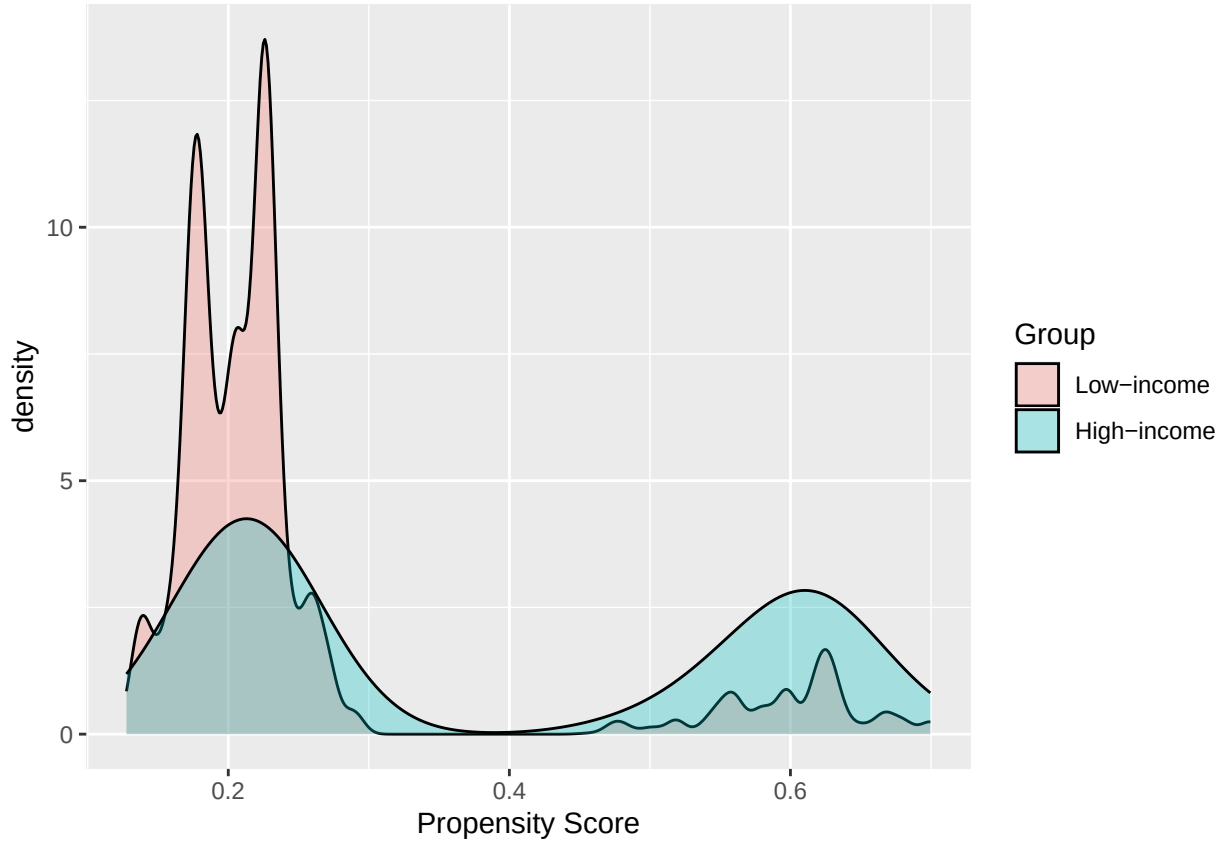
```
# estimate propensity scores using a logit model ####
propensity_model =glm(income~fcollege+urban+wage, data = college,
                      family = binomial(link = "logit"))
tidy(propensity_model)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -2.33     0.251      -9.30 1.38e- 20
## 2 fcollege       1.74     0.0769     22.7  7.89e-114
## 3 urban         -0.321    0.0851     -3.76 1.67e-  4
## 4 wage           0.110    0.0256      4.31 1.64e-  5
```

```
# use logit object to calculate propensity score for all units
college$pscore=predict(propensity_model,type = "response")

# check whether positivity assumption holds
college %>% summarise(pscore_min = min(pscore),
                      pscore_max = max(pscore))
```

```
##   pscore_min pscore_max
## 1     0.1275     0.6995
```

```
# check if there is a systematic difference between the propensity scores of
#treated and controlled units
college$Group=factor(college$income)
ggplot(data = college,aes(x=pscore,fill=Group))+
  geom_density(alpha=0.3) + xlab("Propensity Score") +
  scale_fill_discrete(labels=c("Low-income","High-income"))
```




Here, I first constructed the logit propensity model with three potential confounding covariates:fcollege, urban, and wage. As we can see in the outputted tibble, fcollege and wage exhibit statistically significant(p-value smaller than 0.05), positive correlations with the selection into treatment. Urban, on the other hand, is negatively associated with selection into treatment. Those findings are consistent with the findings in our pre-weighting balance check.

Next, each unit's propensity score(their likelihood of being treated) is computed with the propensity model and their covariates. After computing the propensity scores, I check if the conditional positivity assumption holds in our data. The conditional positivity assumption dictates that $0 < \Pr(D_i = 1 \mid X_i = x) < 1$: given the set of observed covariates, each unit has some likelihood of being selected into the treatment. Since one's propensity score precisely captured the likelihood of being treated given one's covariates, and the values of propensity scores computed with my model are between 0.12 to 0.70, it is safe to claim that this assumption holds for my data.

I also visually compare the propensity score distribution of the two groups. As we can see in the last plot, the low-income group's propensity scores cluster at the range of 0 to 0.3, while having a very low density at the higher propensity score values, if compared to the high-income groups. At the range of propensity score of 0.5 to 0.7, the treatment group has approximately twice of the density of the control group. After confirming this systematic imbalance of selection into treatment between the two groups, we can proceed to construct a weighted sample to adjust for this.

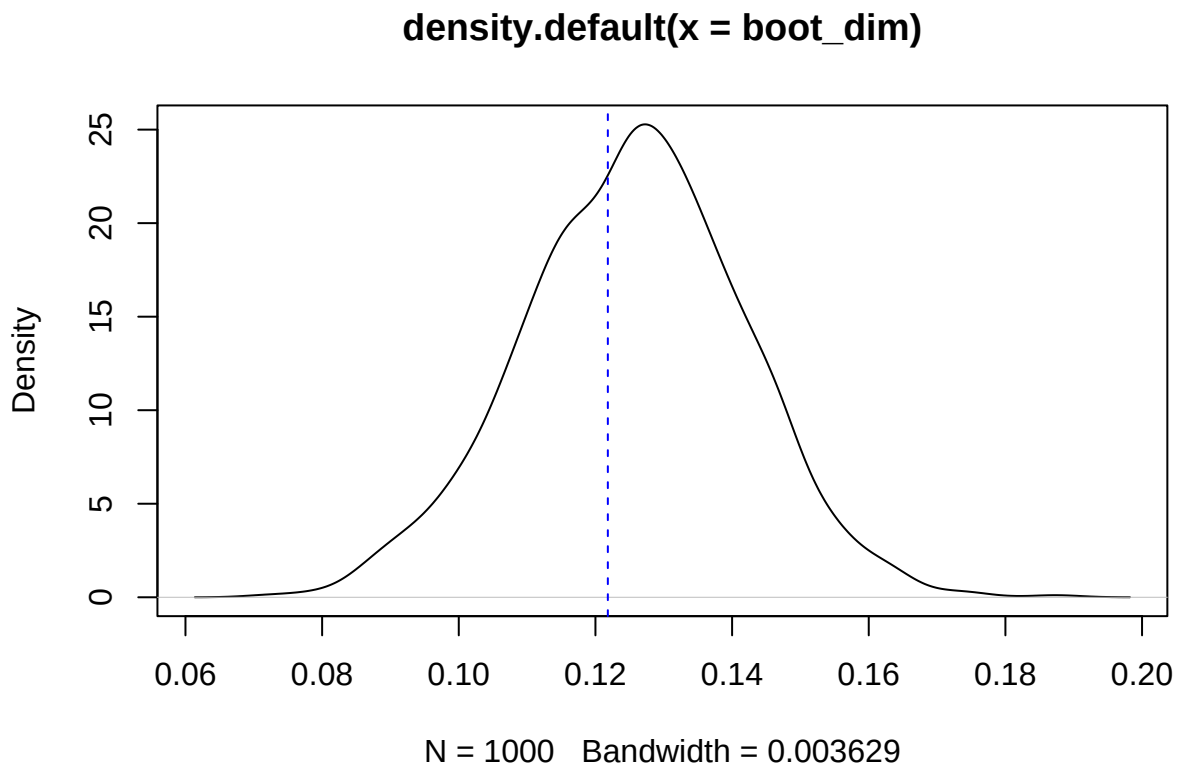
```
#create inverse propensity score weights for treatment and control
college = college %>%
  mutate(weight=ifelse(income==1,1/pscore,1/(1-pscore)))

# run the ipw models and calculate estimate
weighted_diff_in_means=mean(college$weight*college$college*college$income) -
```

```
  mean(college$weight*college$college*(1-college$income))
weighted_diff_in_means
```

```
## [1] 0.1218
```

Then, I proceed to create weights propensity score for all units based on their propensity scores, and calculate the difference between the weighted treated group and the weighted control groups. The point estimate is 0.12. Then, we need to obtain the variance and standard error to compute a 95% confidence interval and see if this estimate is statistically significant.

```
# bootstrap error

boot_dim=c()

for (i in 1:1000) {
  #resample observations
  boot_data=college[sample(1:nrow(college),nrow(college),replace = T),1:9]

  #get propensity scores
  boot_ps=glm(income~fcollege+urban, data = boot_data,
              family = binomial(link = "logit"))

  boot_data$pscore=predict(boot_ps,type="response")

  #weights
   boot_data = boot_data %>% mutate(weight = ifelse(income == 1,
                                                    1 / pscore,
                                                    1 / (1 - pscore)))

  #calculate and sotre difference in means estimate
  boot_dim[i]=mean(boot_data$weight*boot_data$college*boot_data$income)-
    mean(boot_data$weight*boot_data$college*(1-boot_data$income))

}
#plot our bootstrap estimates
plot(density(boot_dim))
abline(v=weighted_diff_in_means,lty=2,col="blue")
```




```
#get standard error estimate and calculate confidence interval
boot_se=sd(boot_dim)
print(boot_se)
```

```
## [1] 0.01616
```

```
weighted_ci=c(weighted_diff_in_means-1.96*boot_se,
              weighted_diff_in_means+1.96*boot_se)
weighted_ci
```

```
## [1] 0.09014 0.15349
```

My bootstrapped 95% confidence interval ranges from 0.09 to 0.15, which does not include zero, indicating a statistcially significant positive relationship between treatment(being from a family of higher income) and outcome(college attendance).

```
balance_post_weighting = college %>% group_by(income) %>% summarise(
  score=weighted.mean(score, weight),
  fcollege_yes=sum(ifelse(fcollege ==1,1,0)*weight/sum(weight)),
  fcollege_no=sum(ifelse(fcollege ==0,1,0)*weight/sum(weight)),
  urban_yes=sum(ifelse(urban==1,1,0)*weight/sum(weight)),
  urban_no=sum(ifelse(urban==0,1,0)*weight/sum(weight)),
  wage=weighted.mean(wage, weight),
  distance=weighted.mean(distance, weight),
```

```
  tuition=weighted.mean(tuition, weight))

print(balance_post_weighting, width = Inf)
```

```
## # A tibble: 2 x 9
##   income score fcollege_yes fcollege_no urban_yes urban_no  wage distance
##    <dbl> <dbl>        <dbl>       <dbl>     <dbl>    <dbl> <dbl>    <dbl>
## 1      0  50.3        0.210       0.790     0.231    0.769  9.52     1.91
## 2      1  52.0        0.210       0.790     0.229    0.771  9.53     1.60
##   tuition
##     <dbl>
## 1   0.810
## 2   0.830
```

Finally, I conduct a post-propensity weighting covariates balance check. Here we can see that previously imbalanced covariates between the two groups are all much closer to each other. The only caveat arise when we examine distance to campus. The distribution of this variable remain relatively stable in our pre and post weighting balance check. This could potentially be indicative of a model misspecification(what if it should be included in the propensity model as a confounder?). On the other hand, it is possible that this is caused by the unit of distance(in 10s of miles), if we convert that to miles, the difference between the two groups is on average 3 miles, which is still a relatively close distance.

Summing up, using an IPW estimator for ATE, I got a point estimate of 0.12 with a 95% confidence interval from 0.09 to 0.15. Compared to the raw estimate I got from difference in means, 0.20, this result accounts for bias potentially introduced by confounders(urban, fcollege, and wage). However, the result needs to be intepreted with caution, since IPW estimator is relying on strong assumption of conditional ignorability, while the model itself is highly sensitive to misspecification.