DS-UA 201: Final Exam

Yiyun (Leo) Yao

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.

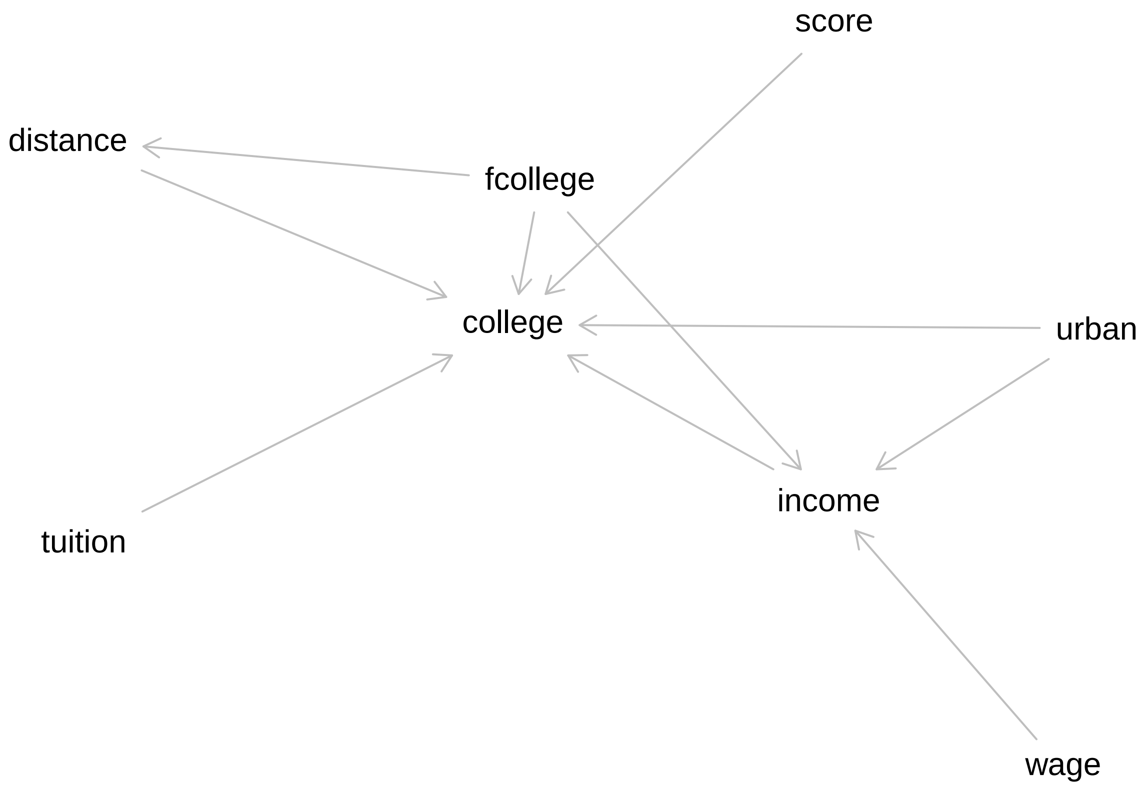
The dataset is `college.csv` and it contains the following variables:

- `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?

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.

```r
library(dagitty)
# Define the DAG using dagitty()
dag <- dagitty("dag {
  college <- income
  college <- tuition
  college <- score
  fcollege -> distance -> college
  income <- wage
  income <- urban -> college
  income <- fcollege -> college
}")
plot(graphLayout(dag))
```



In this DAG, D is the treatment, which is the family income. Y is the outcome, which is whether an individual attends college. In my opinion, whether the individual's father attended college would affect the likelihood of the individual attending college because of colleges' legacy policies. They give better consideration of descendants of college graduates. Whether the individual's father attended college would also affect the distance of the individual from college because the father is more likely to reside near college after graduation. This distance from college impacts the likelihood of the individual attending college because it increases the chances of the individual's participation in pre-college programs. Standardized testing scores affects whether the individual is admitted. Tuition impacts the willingness and affordability of the individual attending college. Minimum wage impacts the family income because the higher the minimum wage is set to be, the higher the family income would be, on average. Whether the family lives in an urban area affects both the family income and whether the individual attends college. If the family lives in an urban area, it is more likely that the family has a higher income than other families who don't live in an urban area. If the family lives in an urban area, the individual is more likely to go to college because urban areas usually have more education resources like community colleges and in general more colleges.

If we want to study the effect of the treatment on the outcome (D on Y), we have to condition on confounders. Confounders are variables that jointly determines both D (the treatment) and Y (the outcome). Confounders create a backdoor path along with the direct path from D (the treatment) to Y (the outcome) and backdoor paths create bias. In this case, whether the father attended college and whether the family attends college are two confounders because they simultaneously affect family income and whether the individual goes to college.

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.

I decide to use ATE rather than CATE or ATT because CATE is suitable for estimating effect of groups with same value for certain covarivate and ATT focuses on the treated groups but we are estimating the effect of the treatment on the outcome overall, in general. So ATE would best satisfy my need in this case without shrinking the sample size as ATT would have done. I decide to use weighting to help estimate the effect. I don't want to use stratification because the two confounders each has two possible values, which, combining with separated control and treatment, gives eight groups, which decreases the sample size. I don't want that to happen because that would cause inaccuracy in the final results. Weighting ensures balance in the covariate distributions of the treated and control samples by reweighting.

The assumptions that need to be satisfied are SUTVA, conditional positivity, and conditional ignorability. To satisfy the SUTVA assumption, we need to have single treatment and make sure there is no spillover effect. Given the treatment is family income, no spillover is ensured because knowing someone's family income does not affect whether one attends college or not. The single treatment assumption is questionable because there are families with income close to the 25,000 line and far from that. These families exist in both lower or higher cases. That causes the units within the treated group as well as the control group to be different. Therefore, we need to address these concerns. Even though it's impossible to make the treatment perfectly "single" in an observation study, conditioning on confounders would solve it to some extent. The conditional positivity assumption holds if the treatment is not deterministic in the confounders, which means the probability of receiving the treatment or control conditionally on the covarivate should be between 0(exclusive) and 1(exclusive). I use weighting for it to be conditionally ignorable. Specifically, I use a logistic regression model to estimate the propensity score for each observation and construct inverse propensity of treatment weights (IPTW) for each observation using the unstabilized weights. The propensity scores indicate the probability of each unit receiving the treatment. I then use them to construct an IPW estimator and report the point estimate for the ATE. Conditional ignorability assumption requires us to assume that the treatment is independent of the potential outcomes conditionally on the covariates. We could satisfy this assumption after weighting since we know that weighted data is as if drawn from a randomized experiment. Therefore, the treatment is independent of the potential outcomes after we pick the confounders to reweight our data.

For the variance estimates, I decide to use bootstrapping because analytical variance estimator are usually difficult to construct, and bootstrapping allows us to capture the uncertainty in our estimator without actually resampling from the population, which is impossible in this case.

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.

```r
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 1.9.3     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 error
```

```
library(haven)
library(estimatr)
```

```
## Warning: package 'estimatr' was built under R version 4.3.2
```

```
library(broom)
```

```
# import the data
df <- read_csv("college.csv", show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```
# data cleaning
df$income <- as.integer(df$income)
df$college <- as.integer(df$college)
df <- df %>% mutate(fcollege = case_when(fcollege == "yes" ~ 1,
                                         fcollege == "no" ~ 0))
df <- df %>% mutate(urban = case_when(urban == "yes" ~ 1,
                                      urban == "no" ~ 0))
```

Check the imbalance:

```
# Standardize the covariates
df_standardized <- df %>%
  mutate(urban_std = urban/sd(urban),
         fcollege_std = fcollege/sd(fcollege))
# Balance between treated and control
balance_table <- df_standardized %>%
  group_by(income) %>%
  summarize(urban_std = mean(urban_std),
            fcollege_std = mean(fcollege_std),
            .groups="keep")
balance_table
```

```
## # A tibble: 2 x 3
## # Groups:   income [2]
##   income urban_std fcollege_std
##    <int>     <dbl>        <dbl>
## 1      0     0.597        0.288
## 2      1     0.438        1.07
```

Both urban and fcollege show imbalance.

```
# Take the absolute differences
abs_balance_diff <- abs(balance_table[1, 2:ncol(balance_table)] -
                          balance_table[2, 2:ncol(balance_table)])
abs_balance_diff
```

```
##   urban_std fcollege_std
## 1    0.1582       0.7807
```

Neither of the absolute differences are 0. So all covariates differ.

```
pscore_model <- glm(income ~ urban + fcollege,
                    data=df_standardized, family=binomial(link="logit"))
tidy(pscore_model)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -1.27     0.0441    -28.9  3.71e-183
## 2 urban         -0.334    0.0850     -3.93 8.65e-  5
## 3 fcollege       1.75     0.0767     22.8  9.93e-115
```

```
# Get the propensity scores for each observation
df_standardized$e <- predict(pscore_model, type = "response")
```

```
# Generate the weights (unstabilized)
df_standardized$wt <- NA
df_standardized$wt[df_standardized$income == 1] <-
  1/df_standardized$e[df_standardized$income==1]
df_standardized$wt[df_standardized$income == 0] <-
  1/(1 - df_standardized$e[df_standardized$income==0])

point_wtd <- mean(df_standardized$wt * df_standardized$college * df_standardized$income
                  - df_standardized$wt * df_standardized$college *
                    (1-df_standardized$income))
point_wtd
```

```
## [1] 0.1261
```

The point estimate here is the average treatment effect of income on college, which is 0.1261, meaning individuals with family income above USD 25,000 per year are 12.61% more likely to attend college.

Then, we use boostrapping to get the standard error and the 95% confidence interval:

```
# Set random seed
set.seed(123)

nBoot <- 1000 # Number of iterations
ate_boot <- rep(NA, nBoot) # Placeholder to store estimates
# For each iteration

for(boot in 1:nBoot){
  # Resample rows with replacement
  college_boot <- df_standardized[sample(1:nrow(df_standardized), nrow(df_standardized),
                                         replace=T),] #replace = T is key!
  # Fit the propensity score model on the bootstrapped data
  college_model_boot <- glm(income ~ urban + fcollege, data=college_boot,
                            family=binomial(link="logit"))
  # Save the propensities
  college_boot$e <- predict(college_model_boot, type = "response")
  # Calculate the weights
  college_boot$wt <- NA
  college_boot$wt[college_boot$income == 1] <-
    1/college_boot$e[college_boot$income==1]
  college_boot$wt[college_boot$income == 0] <-
    1/(1 - college_boot$e[college_boot$income==0])
  # Compute and store the ATE
  ate_boot[boot] <- mean(college_boot$wt * college_boot$college * college_boot$income
                         - college_boot$wt * college_boot$college *
                           (1-college_boot$income))
}

# Take the SD of the ate_boot to get our estimated SE - can do asymptotic inference
sd(ate_boot)
```

```
## [1] 0.01606
```

```
# Asymptotic 95% CI
c(point_wtd - qnorm(.975)*sd(ate_boot),
point_wtd + qnorm(.975)*sd(ate_boot))
```

```
## [1] 0.09464 0.15761
```

Our estimated standard error is 0.01606. The 95% confidence interval is [0.09464, 0.15761]. The 95% confidence interval does not include 0. We could reject the null hypothesis that family income has no effect on whether an individual attends college at $\alpha = .05$. Therefore, we could say that there's convincing evidence that having a higher family income has statistically distinguishable effect in increase the probability for an individual to attend college.