DS-UA 201: Final Exam

Andrew Shao

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(ggdag)
```

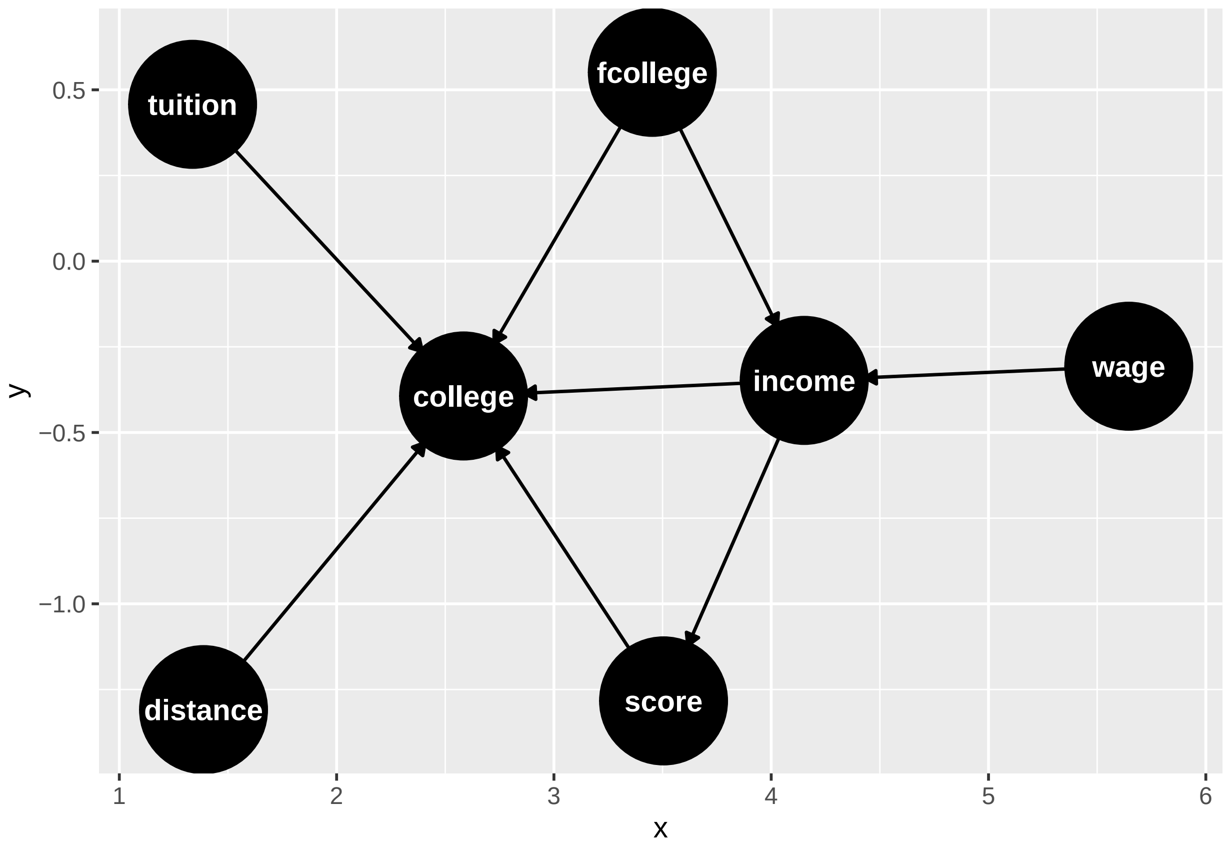
```
## Warning: package 'ggdag' was built under R version 4.3.2

##
## Attaching package: 'ggdag'

## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(dagitty)
library(ggplot2)

dag <- dagify(college ~ income + distance + score + fcollege + tuition,
              score ~ income,
              income ~ fcollege + wage,
              exposure = 'income',
              outcome = 'college')
plot(ggdag(dag, node_size = 22))
```



There are arrows pointing to income from fcollege and wage, which are pretty self-explanatory: incomes are generally higher where wages are higher and having a father who is a college graduate affects income since graduating college unlocks higher paying job opportunities. There is an arrow pointing to score from income because family income affects how much students are able to pay for books/test prep which has a marked effect on students' standardized testing scores. There are arrows pointing to college from tuition, fcollege, income, score, and distance because these are variables which affect students' ability to attend college. These relationships are also pretty intuitive. Income and tuition cost affect whether or not students can afford to pay for college. There is no arrow between the two, however, because income affects financial aid which is used to pay tuition and not directly tuition (to my understanding, I may be wrong). Whether or not your father graduated college affects whether students attend college since it both increases the chances of students getting into college in most cases (legacy admission) as well as likely influences their decision making through parental influence. Score affects it since students must score high enough to get in to college. Distance affects college attendance since accessibility is a consideration students must make when considering college. Confounders are covariates which affect both treatment (income) and outcome (college attendance). The confounder, assuming no unobserved confounders exist, that must be conditioned on is only fcollege in this case since it points to (or affects) both treatment and outcome. On the other hand, you can't condition on score as that is a post-treatment variable which will probably induce bias if conditioned on.

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.

```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 lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks ggdag::filter(), 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
```

```r
library(MatchIt)
```

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

```r
data <- read.csv('college.csv')

m.out0 <- matchit(income ~ score + fcollege + wage + urban + distance + tuition, data =
↪  data, method = NULL, distance = 'glm')
summary(m.out0)
```

```
##
## Call:
## matchit(formula = income ~ score + fcollege + wage + urban +
##     distance + tuition, data = data, method = NULL, distance = "glm")
##
## Summary of Balance for All Data:
##              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance             0.389         0.247           0.699      2.263     0.211
## score               53.329        49.902           0.412      0.926     0.107
## fcollegeno           0.566         0.883          -0.639          .     0.317
## fcollegeyes          0.434         0.117           0.639          .     0.317
## wage                 9.652         9.439           0.155      1.068     0.036
## urbanno              0.815         0.748           0.172          .     0.067
## urbanyes             0.185         0.252          -0.172          .     0.067
## distance.1           1.512         1.920          -0.212      0.629     0.042
## tuition              0.843         0.803           0.116      1.069     0.062
##              eCDF Max
## distance        0.321
## score           0.180
## fcollegeno      0.317
## fcollegeyes     0.317
## wage            0.084
## urbanno         0.067
## urbanyes        0.067
## distance.1      0.097
## tuition         0.111
##
## Sample Sizes:
##           Control Treated
## All          3374    1365
## Matched      3374    1365
## Unmatched       0       0
## Discarded       0       0
```

I will use coarsened exact matching to target the ATE. I chose to target the ATE as opposed to ATT because it is more useful in this case to consider the effect on the entire population as opposed to units that get treated. When considering the effects of family income on the ability of students to attend college, we are probably more concerned with how students of lower income households are negatively impacted as opposed to how higher income students benefit. As a result it doesn't entirely make sense to focus on the effect on treated (higher income) units. Furthermore, higher income families probably share many characteristics with each other which differ from lower income families (difference in treated and control groups), so the ATT may not generalize well. I chose matching because the data is imbalanced between control and treatment; the standardized mean difference for fcollege was 0.639 which is very high. Matching pairs treated units and control units of with similar covariate values, inducing balance. My method of choice is coarsened exact matching since exact matching will drop too many units.

The assumptions that need to be met are SUTVA, positivity, and ignorability. SUTVA is met since there is only a single version of treatment and no spillover. There is a single version of treatment because treatment is just whether or not households make above $25,000 annually, so each unit can't have different values of treatment (you either make or don't make the threshold). Spillover is not an issue since household income in one unit shouldn't affect other units. Presumably by the definition of household, a household only shares income within the same household and not with any other households. Ignorability is satisfied by matching. Units' treatment assignment is now independent of outcome, as units are matched based on similar covariate values, so conditional ignorability is met. After matching, the positivity assumption is met as each unit has a positive probability of receiving treatment since every point in the covariate space has a treatment unit matched.

The choice I will make for variance estimate is robust standard errors. This is allowed because the sample size is large enough (even after matching) to use robust standard error. With a large sample size, bootstrapping can be quite time consuming.

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(estimatr)
```

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

```r
m.out1 <- matchit(income ~ fcollege + wage + urban + distance + tuition, data = data,
↪  method = 'cem', distance = 'glm', estimand = 'ATE')
m.out1
```

```
## A matchit object
##  - method: Coarsened exact matching
##  - number of obs.: 4739 (original), 4419 (matched)
##  - target estimand: ATE
##  - covariates: fcollege, wage, urban, distance, tuition
```

```r
m.data <- match.data(m.out1)

tidy(lm_robust(college ~ income + fcollege + wage + urban + distance + tuition, data =
↪  m.data))
```

```
##          term  estimate std.error statistic   p.value  conf.low conf.high   df
## 1 (Intercept)  0.544836  0.051761   10.5261 1.314e-25  0.443360  0.646313 4412
## 2  incomeTRUE  0.123884  0.016027    7.7298 1.325e-14  0.092463  0.155304 4412
## 3 fcollegeyes  0.238042  0.016397   14.5177 1.099e-46  0.205896  0.270187 4412
## 4        wage  0.004654  0.005601    0.8308 4.061e-01 -0.006328  0.015635 4412
## 5    urbanyes  0.002894  0.017702    0.1635 8.701e-01 -0.031810  0.037598 4412
## 6    distance -0.019141  0.004023   -4.7575 2.023e-06 -0.027029 -0.011253 4412
## 7     tuition -0.035735  0.021658   -1.6500 9.902e-02 -0.078195  0.006726 4412
##   outcome
## 1 college
## 2 college
## 3 college
## 4 college
## 5 college
## 6 college
## 7 college
```

The estimate for effect of having a household income over $25,000 on attending college is 0.123884. The standard error for this estimate is 0.016027 with a 95% confidence interval of [0.092463, 0.155304] and p-value of $1.325 \cdot 10^{-14}$. This result is highly statistically significant, as the confidence interval excludes 0 and the p-value is extremely small. This implies that having a household income higher than the threshold of $25,000 has a strong positive effect on students' ability to attend college, and that wealthier kids are more likely to go to college. This may be concerning if you believe in equal opportunity, and especially if you care about the wealth gap as attending college has a very large impact on your ability to get a higher-paying job which is necessary for socioeconomic mobility (rich kids who have it easier to get into college get richer).