# DS-UA 201: Final Exam

Petra Ivanovic

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.**

Let's first load our data to see what we are working with:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     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 errors
```

```
library(haven)
library(estimatr) # for lm with robust se : ?lm_robust()

# Load  data
data <- read.csv("college.csv")

head(data)
```

```
##   X score fcollege wage urban distance tuition income college
## 1 1 39.15      yes 8.09   yes      0.2  0.8892   TRUE   FALSE
## 2 2 48.87       no 8.09   yes      0.2  0.8892  FALSE   FALSE
## 3 3 48.74       no 8.09   yes      0.2  0.8892  FALSE   FALSE
## 4 4 40.40       no 8.09   yes      0.2  0.8892  FALSE   FALSE
## 5 5 40.48       no 8.09   yes      0.4  0.8892  FALSE    TRUE
## 6 6 54.71       no 8.09   yes      0.4  0.8892  FALSE   FALSE
```

```
str(data)
```

```
## 'data.frame':    4739 obs. of  9 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ score   : num  39.2 48.9 48.7 40.4 40.5 ...
##  $ fcollege: chr  "yes" "no" "no" "no" ...
##  $ wage    : num  8.09 8.09 8.09 8.09 8.09 ...
##  $ urban   : chr  "yes" "yes" "yes" "yes" ...
##  $ distance: num  0.2 0.2 0.2 0.2 0.4 ...
##  $ tuition : num  0.889 0.889 0.889 0.889 0.889 ...
##  $ income  : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ college : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
```

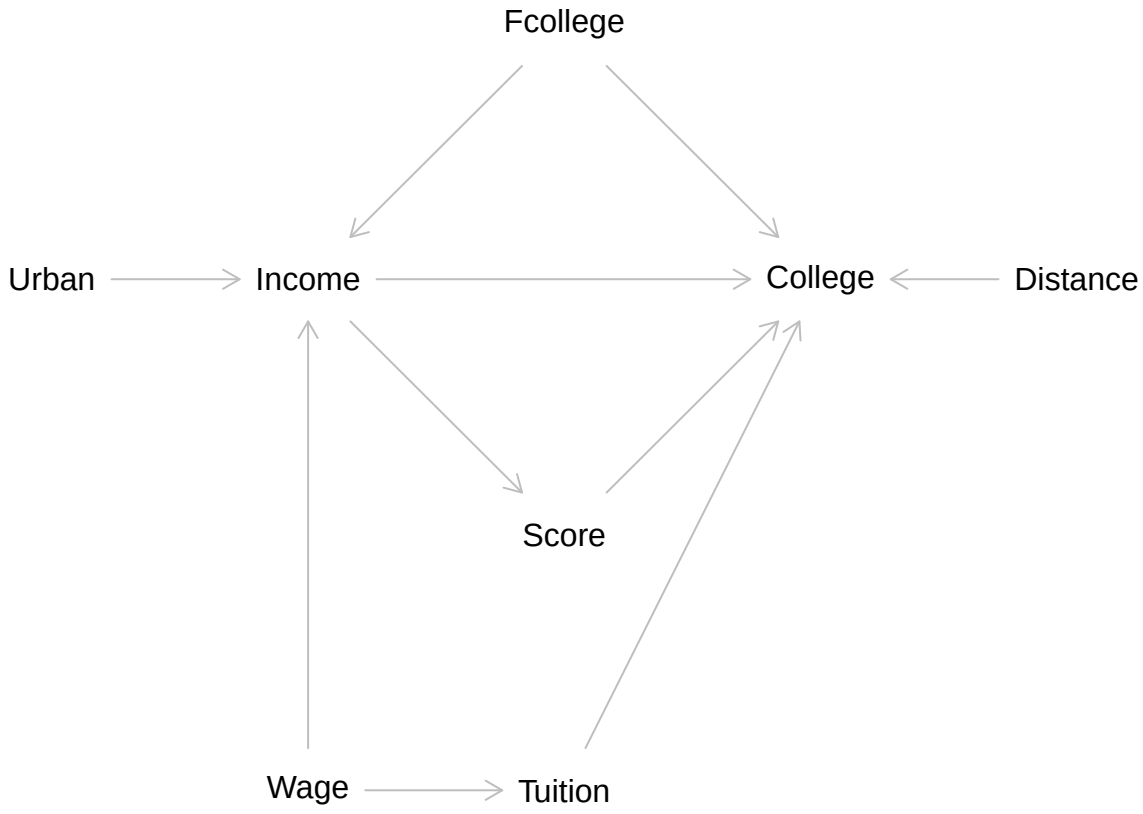
Above we can see the output of data we are working with. We can notice that fcollege, urban, income, college variables are binary but only income and college are in correct TRUE / FALSE format so we might want to consider transforming the other two to also be binary (TRUE / FALSE). We can also see that the rest of our columns are numerical.

Lets design our diagram.

```
library(dagitty)
# ?dagitty(): See the help file

DAG <- dagitty('dag{ Income -> College Fcollege -> Income
               Fcollege -> College Distance -> College Income -> Score
               Score -> College Tuition -> College Wage -> Income Wage -> Tuition
               Urban -> Income}')
coordinates(DAG) <- list(x=c(Income=0, College=2, Fcollege = 1, Score = 1,
                             Distance=3, Tuition = 1, Wage = 0, Urban = -1),
                         y=c(Income=0, College=0, Fcollege = -1, Score = 1,
                             Distance=0, Tuition = 2, Wage = 2, Urban=0))

plot(DAG)
```




Above we can see the DAG (Directed Acyclic Diagram) of all the relationships in our problem. Let's break down all the relationships:

First, let's look at our treatment variable, Income. In this context, Income is a variable that tells us whether the family income is above 25,000$ per year.

Income relationships in the DAG

- to College (I -> C) - this is the relationships we are interested in as we want to see how having family income above 25,000$ (or not) influences whether an individual attended college.
- to Score (I->S) - we can say that family income can have effect on students test scores. Assuming that the tests that were given to high school seniors in the sample are most likely similar to other standarized tests, students coming from wealthier families tend to do better as they have more time and resources to study and prepare (e.g. they do not have to have a job while in school and can afford tutoring). Thus we can say that having higher income can be causing higher test scores.
- from Fcollege (Fc -> I) - we can say that whether the father went to college can influence the family income. This is because, generally, people who do go to college (hopefully) tend to have higher income than those who do not. Thus, having father go to college can influence his income and with that overall income of the family.
- from Wage (W -> I) - we can say that the higher state wage is it is more likely that the family is having an income higher than 25,000$ assuming the wage is good baseline for general economic situation. Note: this also does not have to be true if this state wage is not a good baseline for wages in other fields. (e.g. if it is not growing like others are but less / more quickly.)
- from Urban (U -> I) - we can say that whether a family lives in an urban area can influence whether they get paid more than 25,000$. Specifically, people usually tend to live in more urban areas due to

the ability of finding higher paying positions and more opportunities for work while those in less urban areas usually have less opportunities and with that less pay.

- no direct relationship to Distance - we can say that there is no direct relationship between distance from a college and families income as there are so many different colleges and their locations (urban and rural) so it would be hard to assume that family chose to live somewhere due to the college location as they can't really predict what college their child would get into.
- no direct relationship to Tuition - we can say that one family's income does not have any kind of relationship with college tuition cost (but wage can influence both income and tuition if used as a baseline).

In this part we have covered a lot of our relationships (arrows and lack there of) in our DAG.

Now let's look at the relationship between our Outcome variable, College, that states whether an individual attended college or not.

College relationships in the DAG

- from Income (I -> C) - as mentioned above, this is the relationship we are investigating - whether pay higher than 25,000 influences whether one attends college. We can assume that it does as attending college can be very costly - both in terms of tuition and housing and thus those with higher income have more opportunity to afford it.
- from Distance (D -> C) - we can say that distance from the college can influence whether an individual chooses to go to college. Specifically, if one lives closer to college they might be more likely to go as they can live at home, not have to leave their family (and potentially a job) while living further away comes with additional cost of moving and housing which might prevent them from going.
- from Fcollege (Fc-> C) - we can say that Father being a college graduate makes it more likely an individual will go to college as they will be more encouraged to do so by the father.
- from Tuition (T -> C) - we can say that the tuition can surely influence if someone goes to college or not. More specifically, it tuition is low or free then one have less reasons to not go while if it is costly and they cannot afford it they cannot go to college.
- no direct relationships to Urban - we can say that the Urban does not have a direct relationship as there are really good colleges in urban and rural parts of America and students usually prioritize (or at least they should) how good the college they are going to is rather than where it is located.
- no direct relationships to Wage - we can say that the state hourly wage does not directly influence whether one is going to go to college.

In addition to all the relationships mentioned above, another relationship that we can see in the DAG is from Wage to Tuition (W -> T). This is similar to what we talked about earlier when mentioning the relationship between Wage and Income, where if we consider Wage to be baseline for economic conditions, higher Wage would also make average college Tuition higher as colleges would need more money to pay their employees higher wage (assuming that if Wage variable changes all the wages change in the economy).

Let's now analyse this DAG. We know that a confounder is a variables that influence both our treatment and our outcome and we define colliders as variables that are influenced by both our treatment and outcome. We can see that in our case we do have a confounding variable, Fcollege, as it influences both our treatment (Income) and outcome (College). Thus we should condition on this variable, Fcollege, for the unbias effect of our treatment on our outcome. This is crucial as confounder, Fcollege, can be a common cause for both our treatment, income, and outcome, college attendence, and controlling for the confounder will close the backdoor path of our treatment to our outcome which will help isolate the direct causal effect. We can see that we also have a mediator variable, Score, that is worth pointing out but should not be controled for or conditioned on as it is not a confounder. However, if we wanted to increase the precision (and with that reduce the standard error) of our causal estimate we can control for Distance and Tuition as they have influence on College but they are not confounders so we should not condition on them!

## 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.**

Looking at our problem statement, first thing we can notice is that there is not a lot of information about the problem given. Thus we would want to use a methodology that is very robust to a non-perfect experiment setting. Further, we can assume that we are dealing with an observational study rather than the experiment as it would be very hard (and not to mention unethical) to assign higher and lower incomes to families. This lack of randomization could lead to confounding and with that bias our estimated treatment effect. Thus, we want to use a method that both controls for confounding and is also equipped to deal with observational studies. One such method is Propensity Score Matching. Using Propensity Scores will allow us to reduce bias that is caused by confounding by balancing the covaraiates between treatment and control groups. It is also very useful for us as it is most often used in situations in which random assignment to treatment is not possible (observational studies). This method works by estimating the treatment effect - family income - on our outcome - whether one goes to college - by "matching" people with similar characteristics.

Looking at the best estimand, we could consider both ATE and ATT. Average Treatment Effect, ATE, and Average Treatment Effect on the Treated, ATT, are very similar metrics main difference being that ATE measures the average effect of treatment for the entire population while ATT only measures the effect of treatment on those who have received the treatment. Thus, ATE is a bit broader measure taking into consideration the entire population while ATT only considers those who have received the treatment, in our case having family income over 25,000$ per year. While the measurements are similar, in our case it might be better to look at ATT since we are interested in understanding the effect of having a higher family income on college attendance and it might be easier to gain understaning on this by looking at those who are actually in the higher income group. Further, for practical reasons (especially in propensity score context), ATT would provide a more direct measure of the effect of higher income on college attendance which makes it more actionable for decision/policy making conclusions.

In this study we have multiple assumptions we need to keep in mind. First of those being Conditional Ignorability which assumes that our treatment - family income - is independent of the potential outcomes - attending college - conditional on the covariates. We know that this assumption is valid as so was said in the assignment prompt. Further, our next assumption is Positivity assumption. Positivity means that there is a non-zero probability that each individual will either be a part of treatment or control group. In our case this means that there is a non-zero probability for each family to be earning either below or above 25,000$ which logically always stands (the edge case of 25,000$ can be assigned to either group as per our decision). As always, we also have the assumption of SUTVA. SUTVA means that there is no spillover in our experiment, one units treatment does not influence another unit's outcome, as well as that there is a single version of treatment. In our case there is only 1 unit of treatment as it is families income in $ which will be measured only as below or above 25,000$. Further, there is not a significant spillover with assumptions surveys about participants family incomes were confidential, students will not know each others incomes so that cannot bias their opinion of going to college. Moreover, even if they did know of each others incomes that has little to do with their own ability to attend college as it is only one of the variables we have to consider. While there might be a case of peer influence where one strongly wanting to go to college motivates another or one strongly being against attending college convinces another not to go, it is rarely enough as there are many other variables that have influence on college attendance. Thus, this assumption is also satisfied. Our final assumption is specific for Prosperity Score Matching, the common support assumption. This assumption states that there must be an overlap in the characteristics of the treated and control groups. In other words, we need to be able to find matches in the control group for individuals in the treatment group. Without this overlap (and it being sufficient enough) it can become difficult to make meaningful comparisons.

Finally, we will use robust standard errors to compute our variance estimates. We primarily chose this method

as it is generally the method that is used with Propensity Score Matching. As Propensity Score scores do not rely on many assumptions, robust standard errors compliment the method well as they also provide conservative estimate of uncertainty that also does not rely heavily on the assumptions of homoskedasticity and normality. Further, bootstrapping is generally used in cases where we do not have access to a lot of data. As we have access to data on 4739 individuals, bootstrapping might not be necessary. Finally, calculating robust standard errors is much simpler and faster approach which in the end makes it more efficient and less costly so we should employ it rather than bootstrapping unless bootstrapping is necessary (unless it can lead to much better performance and precision).

## 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.

```
library(MatchIt)

names(data)[names(data) == "distance"] <- "distance_col"

head(data)
```

```
##   X score fcollege wage urban distance_col tuition income college
## 1 1 39.15      yes 8.09   yes          0.2  0.8892   TRUE   FALSE
## 2 2 48.87       no 8.09   yes          0.2  0.8892  FALSE   FALSE
## 3 3 48.74       no 8.09   yes          0.2  0.8892  FALSE   FALSE
## 4 4 40.40       no 8.09   yes          0.2  0.8892  FALSE   FALSE
## 5 5 40.48       no 8.09   yes          0.4  0.8892  FALSE    TRUE
## 6 6 54.71       no 8.09   yes          0.4  0.8892  FALSE   FALSE
```

```
str(data)
```

```
## 'data.frame':    4739 obs. of  9 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ score       : num  39.2 48.9 48.7 40.4 40.5 ...
##  $ fcollege    : chr  "yes" "no" "no" "no" ...
##  $ wage        : num  8.09 8.09 8.09 8.09 8.09 ...
##  $ urban       : chr  "yes" "yes" "yes" "yes" ...
##  $ distance_col: num  0.2 0.2 0.2 0.2 0.4 ...
##  $ tuition     : num  0.889 0.889 0.889 0.889 0.889 ...
##  $ income      : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ college     : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
```

```
# Perform nearest neighbor matching
matchit_model <- matchit(income ~ fcollege + score + wage + urban + distance_col +
                           tuition,
                         data = data,
                         method = "nearest",
                         estimand = "ATT",
                         distance = "glm")
```

```r
# Get the matched data
matched_data <- match.data(matchit_model) # Estimate the ATT on the matched data
outcome_model <- lm_robust(college ~ income, data = matched_data)
#outcome_model

# Extract the coefficient for TRCKNOW and its standard error
att_estimate_nn <- coef(outcome_model)['incomeTRUE']
se_nn <- summary(outcome_model)$coefficients['incomeTRUE', 'Std. Error']

# Calculate the 95% confidence interval
ci_lower_nn <- att_estimate_nn - qnorm(0.975) * se_nn
ci_upper_nn <- att_estimate_nn + qnorm(0.975) * se_nn

# Output the estimate, standard error, and confidence interval
cat("ATT: ", att_estimate_nn, "\n",
    "SE: ", se_nn, "\n",
    "95% CI: [", ci_lower_nn, ", ", ci_upper_nn, "]", sep_nn = "")
```

```
## ATT:  0.1231
##  SE:  0.01744
##  95% CI: [ 0.08889 ,  0.1573 ]
```

Looking at our results we can see that our Average Treatment Effect of those who received the Treatment is 0.1231. This indicates that on average having a family income above 25,000$ increases the chance of attending college by 12.31 percentage points for those in the treated group (those with family incomes above 25,000$). Further we can see that our Standard Error is relatively small at 0.0174 which indicates our ATT estimate is precise. Further our 95% confidence interval [0.08889, 0.1573] does not include zero which indicates that this estimate is statistically significant. Looking at the substantive implications, we can say that (unfortunately) this results seem accurate in the sense that having higher family income makes it more likely an individual attends college. This is partially true due to family influence and partially due to the fact that they can afford it. Due to the fact average cost of college is around 25,000$ per year in USA, having an income above that is practically necessary to even consider attending. This show us that financial stability and family income is indeed a crucial factor in attending college. Further, it shows us that while colleges do offer scholarships, they should make more effort to help families in the lower earning brackets such as those below 25,000$ to make sure education is accessible to all, not just the wealthy. Finally, it is worth noting that this study is from 1986 and is an observational study rather than an experiment and as such might not be as representative of the current times and situation.