ds4e hw1

Task 1

The following statements arise from various theories. In each case, correctly identify the independent variable(s) (there may be more than one) and dependent variable(s) (there may be more than one).

(a) High interest rates reduce consumer spending.

dependent: consumer spending
independent: interest rates

(b) Cycling to work improves health outcomes and reduces congestion.

dependent: health outcomes and congestion
independent: cycling to work

(c) The price of car insurance is mostly related to age, not driving behavior.

dependent: car insurance
independent: age, not driving behavior

(d) In 1850s London, cholera incidence varied by distance to the Broad Street pump.

dependent: cholera incidence
independent: distance to the Broad Street pump

(e) Older people have more retirement savings, and so do college graduates.

dependent: retirement savings
independent: age, college degree

Task 2

Read The New York Times article, “Nike Says Its $250 Running Shoes Will Make You Run Much Faster. What if That’s Actually True?” (nyt running.pdf). Answer the questions that follow.

- **race conditions**
- **weather**
- **gender**
- **age**
- **pre-race training**
- **runner’s previous race times**

(a) What is the treatment of interest in the study?

The treatment of interest here is the Vaporflys shoes, whereas the treatment group is the people who wear the Vaporflys shoes.

(b) Is treatment status randomized (if yes, how; if not, why not)? Is the study an experiment or an observational study?

No. The study itself states that it is not a randomized controlled trial. “Runners choose to wear Vaporflys; they are not randomly assigned them.” However, the article mentions the propensity scores, that aims to control for the likelihood of a person wearing the shoes. This didn’t change the results of the study too.

(c) What is the outcome of interest in the study?

It is the hypothesis that Nike Vaporflys shoes account for an improvement of about 4 percent over runner’s previous time.

(d) What is the implied control group?

People wearing any shoes other than the Nike Vapolflys.

(e) It is difficult to ensure single-blinded treatments in studies such as this. Explain why.

It is pretty difficult to ensure single-blinded treatments in studies such as this. Since runners buy their shoes themselves and they are aware if they are wearing Vaporfly shoes or not. Surely it would be possible to give them fake Vaporfly shoes, had it been an experiment. But the research relies on observational studies.

(f) In this study “weather conditions” are a potential confounder. Explain why — that is,
explain how this confounder might operate.

In good weather, runners may want to put on their good shoes, as they are expensive, and they don't wear them in bad weather, thinking that it might ruin their shoes.

In addition, in good weather, runners feel physically better and stronger, therefore, achieve better results.

By good weather, I mean no rain, no extreme temperatures, but optimal conditions for running.

(g) In the section “Reasons to remain skeptical,” the authors write: “It’s possible that runners wear Vaporflys only when they know they are going to run faster, or that the act of wearing Vaporflys correlates with other things that indicate a runner is going to run faster.” Explain why this is a problem using the appropriate term from class.

This means that there may be a third variable, a possible interference that affects both the change in race results and the change of shoes to Vaperflys. For example, runners’ personal feelings that they are going to run faster. In this case, they will both run faster and wear Vaporflys shoes. Therefore, this would mean that the researchers have established a non-existent connection and advocate the creation of a policy banning the wearing of Vapolfys shoes, which would negatively affect Nike sales.

In addition, there may be Endogeneity, meaning that people wearing Vaporfly shoes and getting good results can cause each other. For example, people wear Vaporfly shoes and run faster because they feel better. Next time they will be able to afford Vaporfly shoes because they performed well at the previous race.

(h) Since the article was written, Nike released an updated shoe named the Vaporfly Next% (the original shoe is named Vaporfly 4%). Suppose researchers want to assess whether the Next% is superior to the 4%. The researchers propose that one group of runners in a marathon be told to wear the Next% while a different group wear the 4%. Such a treatment/control assignment

would be difficult to achieve in practice. Explain the problem using the term from class.

The problem with selecting on the dependent variable. We need some variation both in IV and DV.

For example from a research where we have only two groups of people, one group wearing Next% and a different group wearing 4%; and we conclude that **Next% is superior to the 4%.**

But what if there is a confounder that can’t be detected without having some other variations of other shoes.

Task 3

In 1968, Charles Reep and Bernard Benjamin published a pioneering study on association football (also known as soccer). Among other statistics, Reep and Benjamin reported that approximately 80% of goals scored resulted from sequences of three passes or fewer. To calculate this statistic, Reep and Benjamin took all the goals scored in their data and counted the number of passes leading up to each goal. Some readers concluded that making fewer, more direct passes will increase the probability of scoring a goal.

(a) Identify the independent variable and dependent variable.

Independent variable: number of passes

Dependent variable: probability of scoring a goal

(b) What is wrong with the above conclusion? That is, why is it difficult to draw inferences about causation here? Use the specific term from class, and explain the problem.

Because what Rip and Benjamin found is an association, that is covariation, not a causal relationship. It is easy to establish

an association, just like they did, it is the observation of two things going together. However, establishing a causation requires considering more things, such as counterfactual reasoning. To establish a causal relationship, we need to imagine a world where everything is the same as in this scenario, except for the number of passes. This is quite difficult to do, since we do not have any data on this.

Also, there is no data on other indicators of causality, such as temporal precedence and confounders.

(c) After reading the study, a coach instructs her team to focus on sequences of at most three passes. After implementing this instruction, the team scores more goals. Does this prove the theory that short passing sequences increase the probability of scoring goals? Why or why not?

This still doesn't prove the theory for several reasons:

- This is not based on counterfactual reasoning.
- There may be other counfounders, that is, factors that can lead to both higher scores and shorter passing sequences.
- Endogeneity. Not only can shorter pass sequences lead to higher scores, but higher scores can also lead to shorter pass sequences. In this case, when both DV and IV cause each other, we cannot conclude a causal relationship.

(d) Another coach remains skeptical, asking “What if a team focused on sequences of five or more passes? Would the team score more goals?” What type of reasoning is this? Use the specific term from class.

Counterfactual reasoning, which states that the causal effect of a factor X is a difference between what actually happened and what would have happened had X been different in some way.

Task 4

Associations between variables are straightforward to find—but need not be causal. As we have seen, confounders can cause problems. For each case below, give a plausible confounder for the potential causal relationship claimed. Briefly explain why it might be a confounder.

(a) Young people who hold a passport more often attend college. Obtaining a passport causes college attendance.

This may be related to the socio-economic status of these students. Because people with a higher status could easily get a passport and could afford to attend college. Thus, people with lower socio-economic conditions, even if they have a passport, will be less likely to attend college.

(b) On election day, voters who cast ballots after 5pm tend to vote Democratic. Casting ballots later in the day causes Democratic voting.

People's schedules can be a confounder here. For example, we can say that people who go to school or work full-time are more familiar with literature on democracy and they need democratic conditions in the workplace. Also, since they work full-time, they basically can only vote after 5pm.

(c) Drivers with gold colored cars tend to pay less for car insurance. Choosing a gold car will reduce your car insurance payments.

This may be due to the beliefs of insurance companies that cars of bright colors are less likely to get into car accidents, because they are very noticeable and eye-catching. Thus, even if you paint all your cars in gold, but the insurance company does

not believe that bright colors reduce the likelihood of getting into a car accident, you will not pay less for car insurance.

(d) Students who attended career fairs as sophomores tend to be in full-time employment after graduating. If you want a job after graduating, attend career fairs.

The real factor affecting full-time employment after graduation may be a student's motivation to get a good job, and therefore they both try to attend career fairs and work hard to get a job after graduation. In this case, a student who attends all job fairs, but does not have a strong motivation, is still very likely not to get a job.

Task 5

Read The New York Times article, “Saying No to College” (nyt college.pdf) and answer the questions that follow.

(a) After reading the article, a high school student thinks, “all these people are earning lots of money without a college degree. Surely I can say no to college too!” Which type of bias should the student consider before reaching this conclusion?

Survivor bias.

(b) Explain why the above bias could lead to mistaken conclusions.

It's so easy to focus on people who have dropped out of college and get successful, since we may find them “special". But in this case, we ignore a group of people who also dropped out of school, but did not succeed. Or simply the lack of information about whether people who have not succeeded drop out of college or not leads to mistaken conclusions. Thus this type of thinking leads to biased inferences.

(c) In the article, Mr. Altucher argues that young people can get “ahead in both education and income” by not going to college. What is the causal mechanism by which he suggests this might happen?

He claims that there are thousands of ways to get an education, for example, going to the library, reading a book a day or taking online courses. In addition, since these people who decided to work, instead of going to college, would simultaneously earn some money from their work.

Task 6

The following questions refer to the data provided in the article “The Economic Guide to Picking a College Major” (538 major.pdf) from the American Community Survey. For all questions, you should refer to the first table in the article. Note that while you now have the skills to answer all questions, we have not necessarily given you the explicit code you will need for all questions.

(a) Create an array called myarray that contains numeric data of the median earnings for the top seven ranked college majors (in order from #1 to #7). Please create this array manually (i.e., you do not need to import any data for this homework).

```
import numpy as np
```

```
myarray = np.array([110000, 75000, 73000, 70000, 65000, 65000, 62000])
```

(b) Show that you have indeed created an array by evaluating its type.

```
type(myarray)
```

(c) Use the NumPy package to calculate the mean of myarray and show the result.

```
np. average(myarray)
```

(d) Convert the mean you just generated into an int without simply retyping all of the numbers in the mean you just calculated.

```
int(np.average(myarray))
```

(e) Create a second array called secondarray that contains the median incomes for the majors ranked #8 to #10 from the same table as in question 7(a). As with question 7(a), do this manually (do not import data).

```
secondarray = np.array([[62000,60000,60000]])
```

(f) Use the append command to create a new, third array called fullarray that contains all ten median income numeric observations from rank #1 to rank #10 and show all the values of fullarray.

```
fullarray = np.append(myarray, secondarray)
```

(g) Re-create the first three lines of the table in the article exactly by using a dictionary and then turning it into a pandas DataFrame. (Refer to the first table in the article; you may omit the "$" symbol but you should include a rank column, e.g. 1, 2, 3, and your columns should be numeric where possible).

```
mytable = [{'MAJOR': 'Petroleum Eng.', 'CATEGORY':
'Engineering', 'NUMBER OF MAJORS':
"2,339", "MEDIAN EARNINGS": "110,000"},
{'MAJOR': 'Mining And Mineral Eng.', 'CATEGORY': 'Engineering',
'NUMBER OF MAJORS':
"756","MEDIAN EARNINGS": "75,000"},
{'MAJOR': 'Metallurgical Eng.', 'CATEGORY': 'Engineering',
'NUMBER OF MAJORS':
"856", "MEDIAN EARNINGS": "73,000"}]
```

```
import pandas as pd
df = pd.DataFrame(mytable)
df
```

(h) What is the unit of analysis in this dataset you've just created?

Median Earnings

	MAJOR	CATEGORY	NUMBER OF MAJORS	MEDIAN EARNINGS
1	Petroleum Eng.	Engineering	2,339	$110,000
2	Mining And Mineral Eng.	Engineering	756	75,000
3	Metallurgical Eng.	Engineering	856	73,000
4	Naval Architecture And Marine Eng.	Engineering	1,258	70,000
5	Chemical Eng.	Engineering	32,260	65,000
6	Nuclear Eng.	Engineering	2,573	65,000
7	Actuarial Science	Business	3,777	62,000
8	Astronomy And Astrophysics	Physical Sciences	1,792	62,000
9	Mechanical Eng.	Engineering	91,227	60,000
10	Electrical Eng.	Engineering	81,527	60,000