# Problem Set 4

Answer Key

ECON 480 — Fall 2021

Answers generally go above and beyond what I expect from you. They are meant to show you the correct answer, explain *why* it is correct, and potentially show *several methods* by which you can reach the answer.

## Theory and Concepts

### Question 1

**In your own words, explain the fundamental problem of causal inference.**

---

The fundamental problem of causal inference is that we never know the counterfactuals for the observations that we actually see in the data. A causal effect of a treatment ($\delta_i$) for individual $i$ would be the difference between their outcomes if they were treated ($Y_i^1$) and if they were not-treated ($Y_i^0$), i.e. $\delta_i = Y_i^1 - Y_i^0$. Across many individuals, we could take the average for each group (treated and untreated group). However, in reality, for each individual in our data, we only ever see $Y_i^1$ *or* $Y_i^0$, *never both*. We only see a person's outcome with treatment, or without treatment, so we cannot simply take the difference.

---

### Question 2

**In your own words, explain how properly conducting a randomized controlled trial helps to identify the causal effect of one variable on another.**

---

Randomized controlled experiments are where a pool of subjects representative of a population are randomly assigned into a treatment group (or into one of a number of groups given different levels of a treatment) or into a control group. The treatment group(s) is(are) given the treatment(s), the control group is given nothing (perhaps a placebo, though this is not always necessary), and then the average results of the two groups are compared to measure the true average effect of treatment.

The key is that the assignment must be random, which controls for all factors that potentially determine the outcome (e.g. when measuring individual outcomes, their height, family background, income, race, age, etc). If subjects are randomly assigned, then knowing anything about the individual (e.g. age, height, etc) tells us nothing about whether or not they got the treatment(s). The only thing that separates a member of the treatment group(s) from the control group is whether or not they were assigned to treatment. This ensures that the average person in the treatment group(s) looks like the average person in the control group, and that we are truly comparing apples to apples, rather than apples to oranges.

---

## Question 3

**In your own words, describe what omitted variable bias means. What are the two conditions for a variable to bias OLS estimates if omitted?**

---

All variables that might influence the dependent variable ($Y$) that we do not measure and include in our regression are a part of the error term ($\epsilon$). If we omit a variable ($Z$), it will cause a bias if and only if it meets both of the following conditions: 1. The variable must be a determinant of our dependent variable, $corr(Z, Y) \neq 0$, and thus would appear in the error term, $\epsilon$. 2. The variable must be correlated with one of our independent variables of interest in our regression, $corr(Z, X) \neq 0$.

If both conditions are met, then if we did not include the omitted variable $Z$, our estimate of the causal effect of $X$ on $Y$ would be biased, because our estimate ($\widehat{\beta_1}$) would pick up some of the effect of $Z$. If we include $Z$ as another independent variable, then the $\widehat{\beta_1}$ on $X$ will tell us the precise effect of *only* $X \rightarrow Y$, holding $Z$ constant.

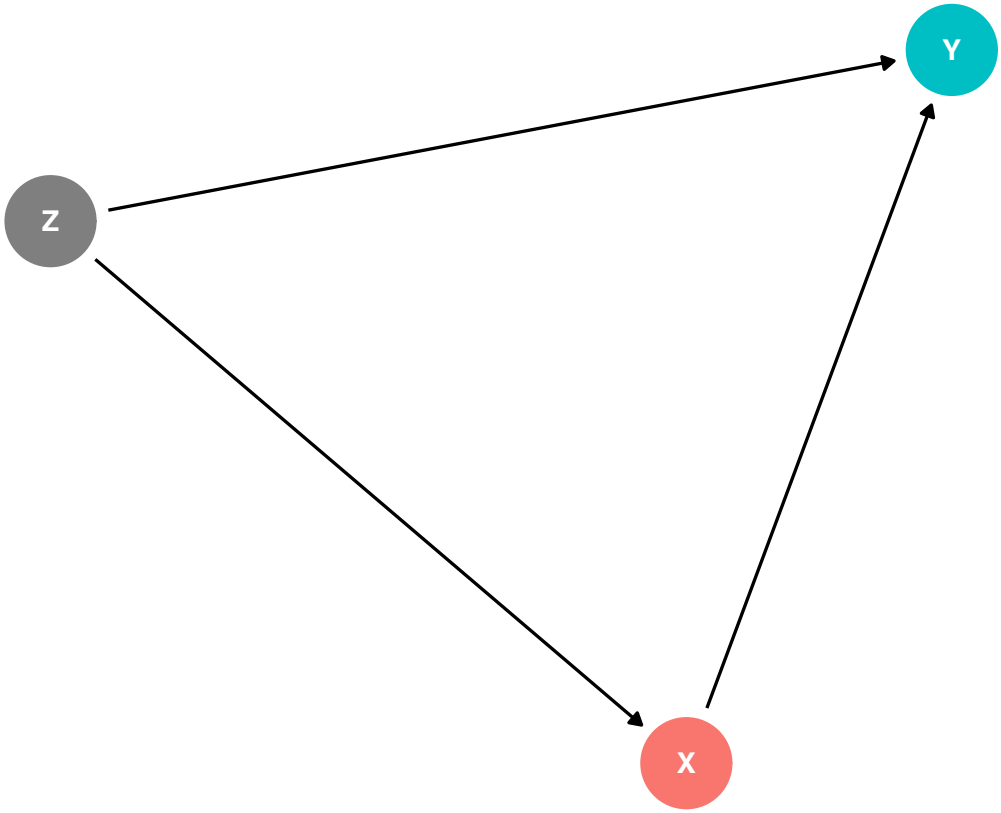
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'ggdag'

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




---

## Question 4

**In your own words, describe what multicollinearity means. What is the cause, and what are the consequences of multicollinearity? How can we measure multicollinearity and its effects? What happens if multicollinearity is *perfect*?**

---

Multicollinearity just means that two regressors (e.g .$X_1$ and $X_2$) are correlated with each other. This fact does *not* bias the OLS estimates of these regressors (e.g. $\hat{\beta}_1$ and $\hat{\beta}_2$). In fact, the reason $X_2$ is included in the regression is because omitting it would cause omitted variable bias, since $corr(X_1, X_2) \neq 0$ and $corr(Y, X_2) \neq 0$. However, the variance of these OLS estimators is increased because it is hard to get a precise measure of $X_1 \rightarrow Y$ because $X_2 \rightarrow Y$ also, and $X_1$ may tend to be certain values (large or small) when $X_2$ is certain values (large or small) so we don't know counterfactuals (e.g. what if $X_1$ were the *opposite* of what it tends to be (large or small) when $X_2$ is large or small).

The strength of multicollinearity is simply given by the value of the correlation coefficient between $X_1$ and $X_2$, $r_{X_1,X_2}$. We can measure the *effect* of multicollinearity on the variance of a regressor ($X_j$)'s coefficient ($\hat{\beta}_j$) with the **Variance Inflation Factor**:

$$VIF = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the $R^2$ from an auxiliary regression of $X_j$ on all of the other regressors.

Multicollinearity is *perfect* when the correlation between $X_1$ and $X_2$ is 1 or -1. This happens when one regressor (e.g. $X_1$) is an exact linear function of another regressor(s) ($e.g. X_1 = \frac{X_2}{100}$. A regression cannot be

run including both variables, as it creates a logical contradiction. In this example, $\hat{\beta}_1$ would be the marginal effect on $Y$ of changing $X_1$ holding $X_2$ constant – but $X_2$ would naturally change as it is a function of $X_1$!

---

### Question 5

**Explain how we use Directed Acyclic Graphs (DAGs) to depict a causal model: what are the two criteria that must hold for identifying a causal effect of $X$ on $Y$? When should we control a variable, and when should we *not* control for a variable?**

---

A Directed Acyclic Graph (DAG) describes a causal model based on making our assumptions about relationships between variables explicit, and in many cases, testable.

Variables are represented as nodes, and causal effects represented as arrows from one node to another (in the direction of the causal effect). We think about the causal effect of $X \to Y$ in *counterfactual* terms: if $X$ had been different, $Y$ would have been different as a response.

When considering the causal effect of $X \to Y$, we must consider *all pathways from $X$ to $Y$* (that do not loop, or go through a variable twice), regardless of the direction of the arrows. The paths will be of two types:

- **Causal (front-door) pathways** where arrows go from $X$ into $Y$ (including through other **mediator** variables)
- **Non-causal (back-door) pathways** where an arrow leads into $X$ (implying $X$ is partially caused by that variable)

Adjusting or controlling for (in a multivariate regression, this means including the variable in the regression) a variable along a pathway closes that pathway.

Variables should be adjusted (controlled for) such that:

1. **Back-door criterion**: no backdoor pathway between $X$ and $Y$ remains open
2. **Front-door criterion**: no frontdoor pathway is closed

The one exception is a **collider** variable, where a variable along a pathway has arrows pointing into it from both directions. This *automatically blocks a path* (whether front door or back door). Controlling for a collider variable *opens* the pathway it is on.

See R Practice on Causality and DAGs for examples.

---

## Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use `R` to *verify* your answers, but you are expected to reach the answers in this section "manually."

### Question 6

A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned the treatment group or into the control group. Half of the people are given the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress the following model:

$$\text{cholesterol level}_i = \beta_0 + \beta_1 \text{dosage level}_i + u_i$$

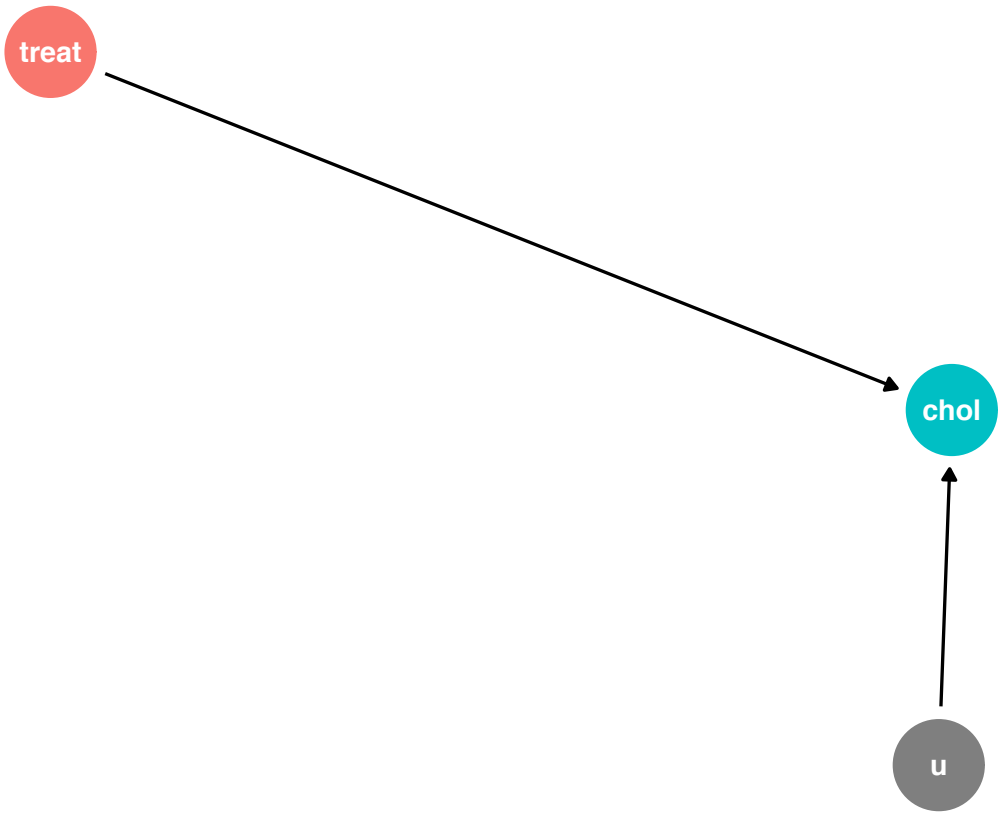
For people in the control group, dosage level$_i$ = 0 and for people in the treatment group, dosage level$_i$ measures milligrams of the active ingredient. In this case, the authors find a large, negative, statistically significant estimate of $\hat{\beta}_1$. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not? Do you expect the estimate to overstate or understate the true relationship between dosage and cholesterol level?

---

Consider the 4th assumption about the error term, $u_i$. Does knowing whether (or how much) a person was treated convey any information about other characteristics that affect cholesterol level (in $u_i$)? Again, we are asking if $E[u|X] = 0$ or $cor(X, u) = 0$.

In this case, the answer is clearly no; knowing whether or not someone received treatment tells us *nothing* else about the person that might affect their cholesterol levels (i.e. age, height, diet, weight, family history, etc., all in $u_i$) because treatment is *randomly* assigned.

In this case, because treatment is exogenous, $E[\hat{\beta}_1] = \beta_1$, $\hat{\beta}_1$ is unbiased.




---

## Question 7

Data were collected from a random sample of 220 home sales from a community in 2017.

$$\widehat{Price} = 119.2 + 0.485\,BDR + 23.4\,Bath + 0.156\,Hsize + 0.002\,Lsize + 0.090\,Age$$

| Variable | Description |
|---|---|
| $Price$ | selling price (in \$1,000s) |
| $BDR$ | number of bedrooms |