Daniel Cabarcas

Assignment 3 – Potential outcomes and OLS
Due date: Thursday, June 11th, 2020 by 5:00pm

DIRECTIONS: The following assignment covers three core parts of the course: potential outcomes, regression and DAGs. Each question is worth 1 point. If you write anything incorrect, you will have points taken off, so be sure that whatever you say it is correct.

Potential outcomes

1. Consider the simple hypothetical example in Table 1. This example involves eleven patients each of whom is infected with coronavirus. There are two treatments: ventilators and bedrest. Table 1 displays each patient's potential outcomes in terms of years of post-treatment survival under each treatment. Larger outcome values correspond to better health outcomes.

Table 1: Perfect doctor example

Patient	Y^1	Y^0	Age	TE	D	Y
1	1	10	29	-9	0	10
2	1	5	35	-4	0	5
3	1	4	19	-3	0	4
4	5	6	45	-1	0	6
5	5	1	65	4	1	5
6	6	7	50	-1	0	7
7	7	8	77	-1	0	8
8	7	10	18	-3	0	10
9	8	2	85	6	1	8
10	9	6	96	3	1	9
11	10	7	77	3	1	10

a. Provide an example of how SUTVA might be violated for treatments of covid-19.

The SUTV assumption requires that all individuals are receiving the same treatment. This might differ if the exposure to the ventilators treatment differs in days of treatment for each patient. If we do not consider contagion dynamics because we have no information on how patients were sampled, we should not be concerned about spillover effects, which is the second requirement. The third assumption relates to the replicability of the results in a larger population. Perhaps the quality of ventilators might differ if a policy attempts to offer universal availability.

b. Calculate each unit's treatment effect (TE).

See table. Procedure -> Individual treatment effect: $\delta_i = Y_i^1 - Y_i^0$

c. What is the average treatment effect for ventilators compared to bedrest? Which type of intervention is more effective on average?

$$ATE = E[\delta_i] = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$$

$$E[Y_i^1] = \frac{(1+1+1+5+5+6+7+7+8+9+10)}{11} = 5{,}45$$

$$E[Y_i^1] = \frac{(10+5+4+6+1+7+8+10+2+6+7)}{11} = 6$$

$$ATE = 5{,}45 - 6 = -0{,}5454$$

According to the ATE, bedrest is more effective on average.

d. Suppose the "perfect doctor" knows each patient's potential outcomes and as a result chooses the best treatment for each patient. If she assigns each patient to the treatment more beneficial for that patient, which patients will receive ventilators and which will receive bedrest? Fill in the remaining missing columns based on what the perfect doctor chooses.

See table. Y is chosen as the highest value between Y^1 and Y^0.

e. Calculate the simple difference in outcomes. How similar is it to the ATE?

$$SDO = E[Y^1|D=1] - E[Y^0|D=0]$$

$$E[Y^1|D=1] = \frac{(5+8+9+10)}{4} = 8$$

$$E[Y^0|D=0] = \frac{(10+5+4+6+7+8+10)}{7} = 7{,}1429$$

$$SDO = 0{,}8571$$

SDO is actually pretty different to ATE both in direction and magnitude.

f. Calculate the ATT and the ATU. How similar are each of these to the SDO? How similar are each of these to the ATE?

$$ATT = E[Y^1|D=1] - E[Y^0|D=1]$$

$$E[Y^1|D=1] = 8$$

$$E[Y^0|D=1] = \frac{(1+2+6+7)}{4}$$

$$ATT = 4$$

$$ATU = E[Y^1|D=0] - E[Y^0|D=0]$$

$$E[Y^1|D=0] = \frac{(1+1+1+5+6+7+7)}{7} = 4$$

$$E[Y^0|D=0] = 7,1429$$
$$ATU = -3,1429$$

g. Show that the SDO is numerically equal to the sum of ATE, selection bias and heterogeneous treatment effects bias. You will need to calculate the ATE, selection bias and heterogenous treatment effects bias, combine them in the appropriate way, and show that their sum is equivalent to the SDO.

$$SDO = \ ATE + Selection\ Bias + (1-\pi)(ATT - ATU)$$
$$Selection\ Bias = E[Y^0|D=1] - E[Y^0|D=0] = -3,1429$$
$$SDO = -0,5454 + (-3,1429) + (0,6363)(4 - (-3,1429))$$
$SDO = 0,8567$ which is similar to the SDO calculated earlier.

OLS

2. The following two questions ask you to estimate two regressions. Report your results in a "beautiful table" labeled Table 1 with a simple description based on parts (a) and (b). You may use this opportunity to learn outreg2 or estout.[1]

a. Create a dataset based on the perfect doctor treatment assignment from part (1). This dataset should *only* contain D, Age and Y. Then estimate the following equation:

$$Y_i = \alpha + \delta D_i + \ \varepsilon_i$$

Report the coefficient on δ . Is it equal to ATE, SDO, ATT or ATU?

b. Now run the following multivariate regression controlling for age.

$$Y_i = \alpha + \delta D_i + \beta Age + \ \varepsilon_i$$

Report the coefficient on δ. Is it equal to ATE, SDO, ATT or ATU? Did controlling for age recover the ATE?

	(1)	(2)
VARIABLES	Model 1	Model 2
Treatment	0.857	0.0142
	(1.430)	(2.340)
Age		0.0202
		(0.0431)
Constant	7.143***	6.355**
	(0.862)	(1.907)
Observations	11	11

[1] I have provided an example for using estout to do this in the /estout subdirectory on github in a file called ols.do, but note that it only creates a LaTeX file. If you want to create something for Word, you will need to use the .rtf format most likely. Read the estout help file online or at Stata.

R-squared	0.038	0.064

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Answering both a. and b., the coefficient on δ (0,857) for Model 1 is equal to the SDO as this is the best estimate of the ATE that can be extracted from the data due to not observing the counterfactuals. The constant in Model 1 is equal to the expectation of the untreated who didn't get the treatment ($E[Y^0|D=0]$). The second model (Model 2) did not recover any of the previous estimates but significantly changed the estimation of the average treatment effect to be much weaker, although neither is statistically significant (but this might be due to the small amount of data).

c. Create a separate table labeled Table 2. This table should have three columns. The first equation is the multivariate regression. The second equation is the auxiliary regression of D onto Age. The third equation regresses Y onto $\widetilde{D}$ which is the residual from the second equation. Compare the coefficient on D from the first equation to the coefficient on $\widetilde{D}$ in the third equation. What does this tell you about how to interpret multivariate regressions?

$$Y_i = \alpha + \delta D_i + \beta Age + \varepsilon_i$$
$$D_i = \beta_0 + \gamma_1 AGE_i + \epsilon_i$$
$$Y_i = \alpha + \delta\, \widetilde{D} + \epsilon_i$$

	(1)	(2)	(3)
VARIABLES	Model 1	Model 2	Model 3
Treatment	0.0142		
	(2.340)		
Age	0.0202	0.0142***	
	(0.0431)	(0.00393)	
D*			0.0142
			(2.280)
Constant	6.355**	-0.403	7.455***
	(1.907)	(0.236)	(0.702)
Observations	11	11	11
R-squared	0.064	0.591	0.000

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The second model shows that the treatment is correlated with age, so that a multivariate model should incorporate these effects. However it turns out that the coefficient for D is basically equal to the coefficient for D*.

Directed acyclical graphs

3. This question is partly based on a 2005 article published in the Journal of Behavioral Medicine that claimed forgiveness improved physical health outcomes.[2]

   Assume that we want to estimate the average causal effect of forgiveness (D) on health (Y) using observational data. Figure 1 represents our belief about how forgiveness and health are related both in the sample and outside the sample.

   We believe that forgiveness (*D*) causes health (*Y*), but we only have data on patients meeting with psychotherapists for mental health treatment (*patients)*.

   Individuals who are more open towards behavioral therapy in the first place (*openness*) become patients (*patients*). We believe these people are also more likely to forgive (*D*).

   Wealth is also important because wealth causes people to see a therapist (*patients*) in part because of their higher willingness to pay for future and present health. Wealth also improves health outcomes. Unfortunately, wealth is not in your data. Wealth is also associated with insurance coverage, which also causes people see therapists (*patients)* and which affects health outcomes.

   And remember – we only have data on patients. Our sample, in other words, consists only of patients.

   a. Write down all backdoor paths between D and Y. Mark whether they are open or closed.

$D \leftarrow P \leftarrow I \rightarrow Y$ opened
$D \leftarrow O \rightarrow P \leftarrow I \rightarrow Y$ closed by collider P
$D \leftarrow O \rightarrow P \leftarrow W \rightarrow Y$ closed by collider P
$D \leftarrow P \leftarrow W \rightarrow Y$ opened

   b. What identification strategy would allow you to estimate the causal effect of forgiveness on health? Assume you aren't limited to merely data on patients.

Identifying causal effects implies closing all backdoor paths. However, based on the backdoors identified, the only way to achieve this would be by conditioning on I, O and W, but O and W are unobserved.

[2] Lawler, et al. (2005), "The Unique Effects of Forgiveness on Health: An Exploration of Pathways", Journal of Behavioral Medicine, vol. 28 (2) April, pp. 157-167.

c. Now assume you only have data on patients. Assume that forgiveness is binary and you calculate the following simple difference in outcomes:

$$Y = \alpha + \delta D + \gamma Insurance + \varepsilon$$

But in this regression, you only use data that you have on patients. Will your estimate of δ identify the ATE? Why/why not? Your answer should indicate whether this control strategy opened up in any backdoors or closed any backdoors.

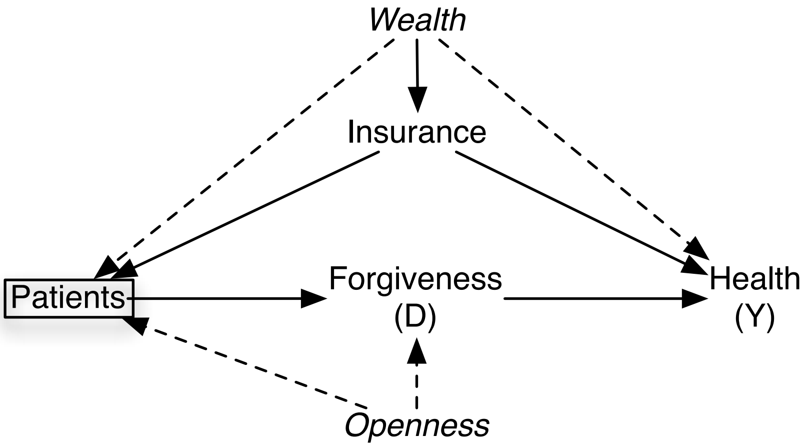

Figure 1: Forgiveness-health study.

With this identification strategy it is not possible to identify the ATE because the second backdoor path identified in a. would be opened. In this case, δ cannot account for the individuals who are not patients, so that counterfactuals cannot be explored even through randomization.

4. Use Figure 2 for the following questions. In all four DAGs (a-d), X is a binary treatment variable and Y is the outcome variable, U and V are unobservable (apologies that they are not dashed lines). S, Z, X and Y are all observable (in your data). For each DAG, answer the following three questions.

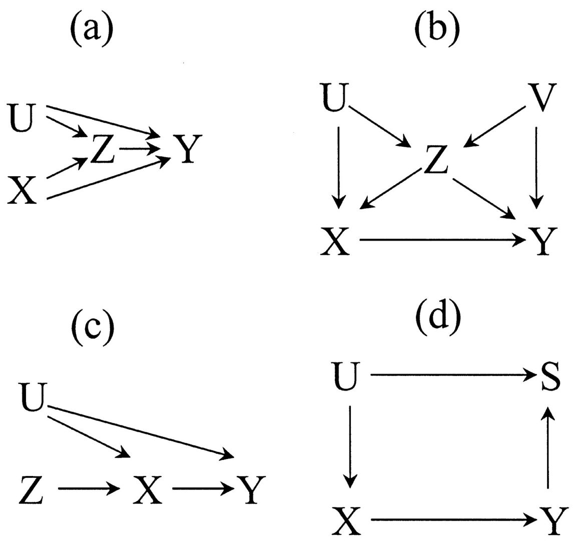

Figure 2: Four DAG examples

a. Write down all backdoor paths from X to Y and indicate whether they are open or closed.
b. Write down a conditioning strategy that satisfies the backdoor criterion. If one does not exist, what is stopping it?

<u>A.</u>

Backdoor paths:

$X \rightarrow Z \rightarrow Y$ opened

$X \rightarrow Z \leftarrow U \rightarrow Y$ closed (because of the collider Z not conditioned on).

Note: I understand the first backdoor path is actually a mediated path, however I noticed, Scott, on the example on gender discrimination and wages that can be found on his book suggests that a mediated path can be a backdoor path (pages 74-75). I use both paths just to be sure but of course the first one cannot satisfy the backdoor criterion.

Conditioning strategy:

If the first path can be considered a backdoor path, then a conditioning strategy would be to use Z as a control. If it is not, then no conditioning strategy is needed as the backdoor criterion is met for the second path. However, controlling by Z would open the second path. So that no condition strategy is available if both are backdoor paths.

<u>B.</u>

Backdoor paths:

$X \leftarrow Z \rightarrow Y$ opened

$X \leftarrow Z \leftarrow V \rightarrow Y$ opened

$X \leftarrow V \rightarrow Z \rightarrow Y$ opened

$X \leftarrow U \rightarrow Z \leftarrow V \rightarrow Y$ closed (because of the collider Z not conditioned on)

Conditioning strategy:

Conditioning on Z would satisfy the first backdoor path, but it is not possible to condition on V to close paths 2 and 3 because it is unobserved, so that no conditioning strategy would satisfy the backdoor criterion.

C.

Backdoor paths:

$X \leftarrow U \rightarrow Y$ opened

Conditioning strategy:

No conditioning strategy is available in order to satisfy the backdoor criterion as the noncollider U cannot be conditioned on since it cannot be observed.

D.

Backdoor paths:

$X \leftarrow U \rightarrow S \leftarrow Y$ closed (collider S not conditioned on)

Conditioning strategy:

No conditioning strategy is need as the backdoor path satisfies the backdoor criterion

DO:

```
clear all
cap log close
set more off
cd "/Users/danielcabarcas/Documents/GitHub/causal-inference-course/Assignment 3"
log using "assignment3", replace

ren var1 patient
ren var2 y1
ren var3 y0
ren var4 age
ren var5 treatment_effect
ren var6 d
ren var7 outcome

label var patient "Patient"
label var y1 "Y1"
label var y0 "Y0"
label var age "Age"
label var treatment_effect "TE"
label var d "Treatment"
label var outcome "Outcome"

drop treatment_effect y1 y0 patient

*a
reg outcome d
//result equals SDO
outreg2 using assignment3_table1.doc, replace ctitle("Model 1") label

*b
reg outcome d age
outreg2 using assignment3_table1.doc, append ctitle ("Model 2") label

*c
reg outcome d age
outreg2 using assignment3_table2.doc, replace ctitle("Model 1") label

reg d age
outreg2 using assignment3_table2.doc, append ctitle("Model 2") label

predict resid, residuals
label var resid "D*"

reg outcome resid
outreg2 using assignment3_table2.doc, append ctitle("Model 3") label

log close
```