Exercise: Controlling for Confounding

Let's simulate some fake data and see when we are able to recover the correct treatment effect just by controlling for covariates.

1. First, let's generate two potential confounder variables for 10,000 people.

(a) The variable 'age' should be drawn randomly from the normal distribution with mean 40 and standard deviation 7.
(b) The variable 'gender' should be drawn randomly from the binomial distribution with a 0.5 probability of being male or female. (Let's assume gender=1 is female).

```
set.seed(54321)
N <- 10000
d <- tibble(age=rnorm(N,40,7),
            gender=rbinom(N,1,0.5))
```

2. Our outcome is going to be attitudes to redistribution. Use the expressions below to simulate potential outcomes, with a constant treatment effect of 5.

$$y_0 = N(10,2) + \frac{age}{5} - 12 * gender$$

$$y_1 = y_0 + 5$$

```
d <- d %>% mutate(y_0=rnorm(N,10,2) + age/5 - 12*gender,
                  y_1=y_0+5)
```

3. Treatment D is receiving a government social program, but treatment is **not** randomly assigned in any way. Instead, treatment depends on age and gender, with older women more likely to be treated. Binary (1/0) treatment is determined by the following expression:

$$D = \begin{cases} 1 \text{ if } (20 * gender + age) > 50 \\ 0 \text{ else} \end{cases}$$

```
d <- d %>% mutate(D=case_when(20*gender+age>50~1,
                              T~0))
```

4. Calculate observed outcomes based on potential outcomes and treatment.

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

5. Let's start with the 'naive' regression of the outcome on treatment with no controls. How does the result compare to our assumed treatment effect?

```
d %>% lm(y_obs~D, data=.) %>% stargazer(title="Q5",
                                        header=F)
```

6. Is the result of the 'naive' regression larger or smaller than it should be? Why? Be specific and be concrete about the type and direction of bias.

Table 1: Q5

	Dependent variable:
	y_obs
D	-4.291^{***}
	(0.091)
Constant	16.590^{***}
	(0.064)
Observations	10,000
R^2	0.184
Adjusted R^2	0.183
Residual Std. Error	4.526 (df = 9998)
F Statistic	$2{,}247.358^{***}$ (df = 1; 9998)
Note:	*p<0.1; **p<0.05; ***p<0.01

The result is smaller than we specified. Gender is a confounder: Women are more likely to be treated but have much lower potential outcomes, so the y_1s we see are more often for women and lower than for the overall sample, while the y_0s are higher.

7. Draw (by hand and take a photo) the causal diagram (DAG) associated with the variables and equations we have generated above.
8. Identify all the back-door paths connecting treatment (D) and the outcome (Y). To **block** these back-door paths, which variables do we need to control for?

Gender and age.

9. Run a regression that controls for the variables you identified in Q8. Compare the result to the 'correct' treatment effect we assumed at the start.

```
d %>% lm(y_obs~D + gender + age, data=.) %>% stargazer(title="Q9",
                                                        header=F)
```

10. That was the easy part. Now let's make our example more complicated. First, imagine that attitudes to redistribution are **NOT** our outcome variable of interest. Instead, our outcome of interest is how much people gave to charity. Generate a new variable for charitable donations which directly depends on attitudes to redistribution (plus some noise):

$$Charity = N(0, 2) + 2 * y_{obs}$$

Note that the treatment effect of D on $Charity$ should now be 10: D increases y_{obs} by 5, and then this effect is multiplied by 2 to create $Charity$.

```
d <- d %>% mutate(Charity=rnorm(N,0,2) + 2*y_obs)
```

11. Draw the causal diagram (DAG) for this new set of variables including $Charity$.
12. Now charity outcomes depend on treatment, age, gender and attitude to redistribution (y_{obs}) so let's regress charity on all of these variables. Is the estimate of the treatment effect correct? What's wrong with this regression?

```
d %>% lm(Charity~D + gender + age + y_obs, data=.) %>% stargazer(title="Q12",
                                                                  header=F)
```

Table 2: Q9

	Dependent variable:
	y_obs
D	5.077***
	(0.090)
gender	−12.076***
	(0.086)
age	0.202***
	(0.003)
Constant	9.941***
	(0.136)
Observations	10,000
R^2	0.838
Adjusted R^2	0.838
Residual Std. Error	2.018 (df = 9996)
F Statistic	17,199.540*** (df = 3; 9996)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Q12

	Dependent variable:
	Charity
D	−0.206**
	(0.104)
gender	0.247*
	(0.149)
age	−0.002
	(0.004)
y_obs	2.007***
	(0.010)
Constant	−0.044
	(0.170)
Observations	10,000
R^2	0.960
Adjusted R^2	0.960
Residual Std. Error	2.031 (df = 9995)
F Statistic	60,531.670*** (df = 4; 9995)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimate is less than expected because we are controlling for a post-treatment variable, which is removing part of the treatment effect.

13. Propose and implement a regression to correctly estimate the effect of the treatment on *Charity*, only including appropriate controls.

```
d %>% lm(Charity~D + gender + age, data=.) %>% stargazer(title="Q13",
                                                     header=F)
```

Table 4: Q13

	Dependent variable:
	Charity
D	9.983***
	(0.202)
gender	−23.988***
	(0.193)
age	0.403***
	(0.008)
Constant	19.907***
	(0.306)
Observations	10,000
R^2	0.803
Adjusted R^2	0.803
Residual Std. Error	4.531 (df = 9996)
F Statistic	13,558.740*** (df = 3; 9996)
Note:	*p<0.1; **p<0.05; ***p<0.01

14. Finally, let's forget about the *Charity* variable and create two more variables, *Education* and *Religion*, as described below. Then redefine potential outcomes and recalculate observed outcomes, where the outcome is Attitude to Redistribution.

$$education = N(10, 1)$$

$$religion = \begin{cases} 1 \text{ if } (2 * age + 2 * education + N(3, 1)) > 100 \\ 0 \text{ else} \end{cases}$$

$$D = \begin{cases} 1 \text{ if } (20 * gender + age) > 50 \\ 0 \text{ else} \end{cases}$$

$$y_0 = N(10, 2) + \frac{age}{5} - 3 * gender + 5 * education$$

$$y_1 = y_0 + 5$$

```r
d <- d %>% mutate(education=rnorm(N,10,1),
                  religion=case_when(2*age+2*education+rnorm(N,3,1)>100~1,
                                     T~0),
                  y_0=rnorm(N,10,2) + age/5 - 3*gender + 5*education,
                  y_1=y_0+5,
                  y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

15. Draw the causal diagram (DAG) including all of these variables. Identify all the back-door paths. How many are there? Which ones are open and which ones are blocked?

The key point is that the path running through religion is a collider and so is already blocked.

16. Identify a minimum set of variables you need to control for to recover the correct treatment effect (don't run this regression yet).

Gender and age are sufficient.

17. If we just received a dataset with our treatment, outcome, gender, age, and religion variables, we might consider these last three all potential confounder variables and want to control for them. Run the regression of observed attitude outcomes on treatment, gender, age and religion. Is the estimated treatment effect correct? Why/why not?

```r
d %>% lm(y_obs~D + gender + age + religion, data=.) %>% stargazer(title="Q17",
                                                                  header=F)
```

Table 5: Q17

	Dependent variable:
	y_obs
D	5.976***
	(0.246)
gender	−3.813***
	(0.234)
age	0.010
	(0.014)
religion	3.002***
	(0.181)
Constant	65.780***
	(0.493)
Observations	10,000
R^2	0.193
Adjusted R^2	0.193
Residual Std. Error	5.349 (df = 9995)
F Statistic	597.563*** (df = 4; 9995)
Note:	*p<0.1; **p<0.05; ***p<0.01

This estimate is inaccurate because we are controlling for a collider (religion).

18. Now remove religion from your regression. Is the estimated treatment effect correct?

```r
d %>% lm(y_obs~D + gender + age, data=.) %>% stargazer(title="Q18",
                                                       header=F)
```

Table 6: Q18

	Dependent variable:
	y_obs
D	4.972***
	(0.242)
gender	−2.960***
	(0.231)
age	0.196***
	(0.009)
Constant	60.188***
	(0.366)
Observations	10,000
R^2	0.171
Adjusted R^2	0.170
Residual Std. Error	5.422 (df = 9996)
F Statistic	685.707*** (df = 3; 9996)
Note:	*p<0.1; **p<0.05; ***p<0.01