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.

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

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

$$y_1 = y_0 + 5$$

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. Binary (1/0) treatment is determined by the following expression:

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

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

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?

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

7. Draw (by hand) 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?

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.

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 $Charity$.

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?

13. Now run the regression of Charity on just treatment, age and gender, i.e. excluding attitude to redistribution y_{obs}. Is the estimate of the treatment effect correct?

14. Finally, create two more variables, Education and Religion. Then redefine potential outcomes and recalculate observed outcomes.

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

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

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

$$y_1 = y_0 + 5$$

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

16. Identify a minimum set of variables you need to control for to recover the correct treatment effect?

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. Run the regression of observed attitude outcomes on treatment, gender, age and religion. Is the etimated treatment effect correct? Why/why not?

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