PhD Econometrics (ECON50580)
Problem Set 1: Causality

This problem set will be graded. Rules:

- You have to work in groups of 3-5 students
- Submit your solutions in 1 pdf
- The code should be in the appendix. (If you use R Markdown, please no code chunks; separate answers and code)
- Results should be presented graphically or in tables. No screenshots from statistical software.
- Submit via Brighspace; it is enough if one member of the group submits.
- Use version control for the empirical exercises, and show evidence thereof (e.g. a screenshot of git)

Submission deadline: Monday, January 23, 11:59:59pm.

1 Theory I: DAGs and Potential Outcomes

a) Consider the following threshold model with a binary treatment D, an additional (binary) covariate x, the outcome y and an i.i.d error term with mean zero ε_i

$$\begin{aligned} y_i &= \alpha + \beta D_i + \gamma x_i + \rho \varepsilon_i \\ x_i &= \mathbf{1}(\kappa + \delta D_i + \pi \varepsilon_i > c). \end{aligned} \tag{1}$$

The second part of Equation 2 means that $x_i = 1$ if the sum on the right-hand side is greater than a threshold c and $x_i = 0$ otherwise. Assuming that δ=0, show that the estimator of the average treatment effect (ATE), $\Delta = E(\boldsymbol{y}|D = 1, x = X) - E(\boldsymbol{y}|D = 0, x = X)$ equals $\beta + bias$, characterise the bias (i.e. derive the exact formula) and explain what the bias means.

b) Now assume that $\delta \neq 0$. Write down a DAG that represents the model. Should the researcher account for x to deconfound the treatment effect?

c) Now assume that $\delta \neq 0$ and $\gamma = 0$. Using the potential outcomes framework, show that the bias $E(\boldsymbol{y}|D = 1, x = X) - E(\boldsymbol{y}|D = 0, x = X) - \beta$ can be re-written as a weighted average of $E(\varepsilon|D = 1, x = X) - E(\varepsilon|D = 0, x = X)$ across groups with $x = 1$ and $x = 0$. Provide a brief interpretation of the bias term.

d) Suppose $\beta > 0$. Derive conditions under which the inclusion of x would lead to the under-estimation of β, i.e. $E(\widehat{\beta}) < \beta$.

2 Theory and Simulation

2.1 The Gender Pay Gap

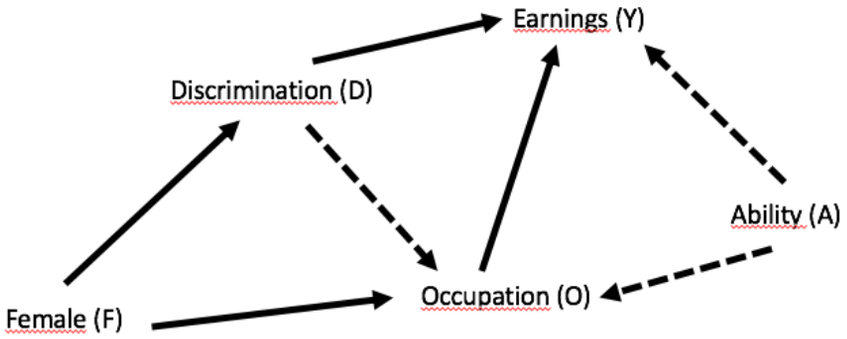


FIGURE 1 – DIRECTED ACYCLICAL GRAPH: GENDER DISCRIMINATION AND EARNINGS

A researcher wants to estimate the effect of gender discrimination on earnings. To deconfound the causal effect, he/she develops the causal diagram shown in Figure 1. All variables except A are observed in the dataset. F is a dummy variable that equals unity if a person is female. O is a set of dummy variables for broad occupational categories. D is a dummy variable indicating if a person is being discriminated or not. Assume that only women are discriminated against. The arrow from D to O is dashed because it is theoretically unclear whehter we should expect an effect of discrimination on occupation.

a) Provide an intuitive explanation for each arrow in Figure 1. Provide an explanation for the absence of an arrow between F and Y. In your view, should there be additional arrows and/or variables in the diagram? If so, which ones?

b) Write out the paths from D to Y. Now assume that you observe ability and can control for it. Explain whether it makes sense (or not) to additionally control for the following: i) only F; ii) only O; iii) both.

c) Assume that A is unobservable. Explain why controlling for O can lead to collider bias.

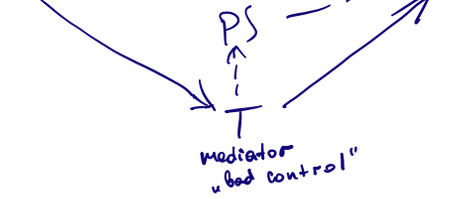
d) Illustrate the collider problem in a simulation based on the above causal diagram. To do so, create a (simulated) dataset that represents all the arrows in Figure 1. It is sufficient to approximate O with one dummy variable. You will have to run several regressions; at the least, show the following regressions: i) Y on D; ii) Y on D controlling for O; iii) Y on D controlling for O and A. Run further regressions if needed and explain why the coefficients differ (or not). The task is here to show convincingly that controlling for variables on the causal path can lead to collider bias. This is what researchers would do in methodological papers, conference discussions or referee reports.

2.2 Application: Miscarriages and the Outcomes of Subsequent Children

In a new paper, the authors study the effect of a mother having a miscarriage on the outcomes of subsequent children. Prior research has shown that miscarriages, while common, can have traumatic effects on women and, by extension, on families. It is thus plausible that a miscarriage affects the outcomes of children that were subsequently conceived and born, for example through changes in parenting styles.

The authors undertake several steps towards establishing causality. They refer to a large number of studies showing that the likelihood of having miscarriages appear to be unrelated to mother or family characteristics, and provide balancing tests in support of this assumption. A second challenge is to find a suitable control group. They restrict the sample to families with two children; the treatment group had a miscarriage in between the births of both children, the control group had no miscarriage.

a) While this identification strategy appears plausible at first, the choice of control group may induce a bad control problem (i.e. the choice is equivalent to conditioning on a mediator). Construct



a DAG to explain where the bad control problem could lie and why this might bias the estimates (Hint: it has to do with the decision to have another child after a miscarriage). Discuss under what conditions the assumption that having a miscarriage is random is sufficient for establishing causality.

b) Using potential outcomes notation, derive the bias in the estimation of the ATE that results from the bad control problem found in a). Explain in what direction the bias could likely go.

c) Propose tests that could potentially show that the bias is quantitatively unimportant (after all, no identification strategy is perfect; so showing that a bias does not matter is often what is needed). Please be brief here.