DXE_EMTR 2023
Second assignment (20% of total grade)

Please submit the assignment by 28 Nov in the IS MUNI system. You are allowed and encouraged to work in groups of maximum size 3. Please don't forget to submit your R-code too.

1 Causal graphical models (3% + 1% bonus)

[This exercise will motivate you to try out `dagitty` *- a powerful tool for analyzing causal models.]*

We are interested in exploring the causal association between teacher's teaching and students' labor market outcomes. Assume the following causal structure:

Outcome Y (some future labor market outcome such as employment or earnings) is assumed to be a function of:

- observed teacher's performance X_1 (e.g. clarity of exposition, availability... - measured by a survey),
- observed student characteristics X_2 (e.g. past exam results),
- unobserved student characteristics U (e.g. motivation, extra curricular activities, grit...).

Teacher's performance X_1 is assumed to be a function of

- student characteristics X_2 (e.g. students have an effect on teachers' performance, highly selective students demand more from their teacher),
- some unobserved variables V,

(Observed) student characteristics X_2 is assumed to be a function of

- unobserved student characteristics U.

We are interested in the causal effect of X_1 on Y.

- Plot the causal graph (denote it as G).
- List all the causal and non-causal paths from teacher's characteristics (X_1) to wages (Y).
- Determine if it is possible to identify the causal effect of X_1 on Y based on observed probability distribution of (X_1, X_2, Y).
- If it is possible, state the minimum adjustment set and propose an estimator. If it is not possible explain in detail why.
- (Bonus +1%) Complement your analysis of G with a small simulation study. Convince the reader that the estimator that you proposed is able to recover the true (total) average treatment effect of discrimination on wages. Make any reasonable simplifications in order to illustrate your point.

You may find `dagitty` helpful for these tasks (either in `R` or at `www.dagitty.net`).

2 Selection on Observables (10%)

[This exercise is meant to boost your understanding of the different concepts in the treatment effects/policy evaluation field.]

Consider the following *hypothetical situation*. We have a population (not a sample) of 20 individuals and we happen to know both their counterfactual earnings: if they go to job training programme ($D = 1$) they will receive $Y(1)$. Or if they do not participate in the job training programme ($D = 0$) they will get $Y(0)$.

This programme was randomly offered to some individuals ($Z = 1$), while others were not offered this programme ($Z = 0$).

Suppose that, in this hypothetical scenario, we also happen to know contrafactual training participation $D(Z)$, so that we know $D(1)$ (for $Z = 1$) and $D(0)$ (for $Z = 0$) for every individual.

We also have information on the different background they have: some individuals have higher-education ($X = 1$) while others do not ($X = 0$).

Unit	$Y(1)$	$Y(0)$	$D(1)$	$D(0)$	Z	X
1	20	19	1	1	0	1
2	30	28	0	1	0	0
3	19	20	1	1	0	0
4	17	14	1	0	0	1
5	28	22	1	0	0	1
6	22	22	1	0	0	0
7	29	27	1	1	0	0
8	22	23	1	0	0	1
9	21	22	1	1	0	1
10	21	19	1	1	0	1
11	25	19	0	0	1	0
12	24	24	1	1	1	1
13	21	20	1	1	1	0
14	25	22	0	0	1	0
15	10	11	1	1	1	1
16	35	23	1	1	1	1
17	22	20	1	0	1	0
18	31	30	1	0	1	1
19	21	20	0	1	1	0
20	33	25	1	1	1	1

(a) What is the true average treatment effect (ATE) of job training programme on earnings?
(b) What is the true average treatment effect on the treated (ATT) of job training programme on earnings?
(c) What is the true average treatment effect on the untreated (ATU) of job training programme on earnings?
(d) What is the proportion of always-takers, compliers, defiers, never-takers (in a relationship $Z \to D$)?
(e) What is the local average treatment effect (the true average treatment effect on the compliers)?
(f) What is the true average treatment effect on the treated of job training programme on earnings for those with higher education?
(g) What is the true average treatment effect on the treated of job training programme on earnings for those without higher education?
(h) Is the selection-on-observables a reasonable assumption here?
(i) Is the instrument Z relevant for D?
(j) Does the monotonicity condition hold?
(k) Plot the distribution of individual treatment effects.

This information is obviously not available to an outside analyst, who only observes (Y, D, Z, X), where $Y = Y(1) \cdot D + (1 - D) \cdot Y(0)$ is the observed earnings and $D = D(1) \cdot Z + (1 - Z) \cdot D(0)$ is the observed programme participations.

(l) Using the observed data (Y, D, Z, X) (reconstructed from the hypothetical distribution of $(Y(1), Y(0), D(1), D(0), Z, X)$) calculate unadjusted differences in mean observed outcomes job training participants ($D = 1$) and non-participants ($D = 0$), that is $E[Y|D = 1] - E[Y|D = 0]$.

Now assume that the 20 observations we have is only a sample from some larger population.

(m) Compare the estimated LATE (that assumes monotonicity) with the result from (e). Explain your findings.

(n) Calculate the average treatment effect on the treated (ATT) based on matching on observable variable X - that is 2 groups.

You can use you favourite software to perform these calculations, but you don't have to. If you will use `R`, this may be helpful:

```
df <- data.frame(
  Y1 = c(20, 30, 19, 17, 28, 22, 29, 22, 21, 21, 25, 24, 21, 25, 10, 35, 22, 31, 21, 33),
  Y0 = c(19, 28, 20, 14, 22, 22, 27, 23, 22, 19, 19, 24, 20, 22, 11, 23, 20, 30, 20, 25),
  D1 = c(1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1),
  D0 = c(1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1),
  Z = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
  X = c(1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1)
)
```

3 Replication (7% + 1% bonus)

[The purpose of this exercise is to try a replication - it is very likely that replicating some existing results from published papers during your PhD path will be directly useful for your research.]

Here are data archives of Joshua Angrist `https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive` and of Daron Acemoglu `https://economics.mit.edu/faculty/acemoglu/data`. There are datasets and replication files in STATA (another program that many economists use for conducting statistical analyses).[1]

- Choose one paper and read it. Preferably choose a paper that is close to your research agenda, if this is possible. Feel free to choose a paper that is not at Angrist's nor Acemoglu's website if that is more appropriate for your research.
- You may want to choose a paper that is not too complicated.
- Replicate the main results in `R`. You don't have to replicate all the results (e.g. those in appendices etc) in the paper, only the most interesting set of main results. R-package `haven` may be useful for you for loading the datasets into `R`.
- Explore if/how these results are sensitive to the functional form specification of the regressions or other model choices. These are some examples you may/may not consider:
  - check whether adding a quadratic term of certain covariate into a regression changes the results substantially,
  - check whether adding a relevant interaction term changes the results substantially,
  - whether having the outcome variable in logarithm (or any other sensible transformation - depending on the empirical context) leads to very different conclusions,
  - look if the results hold if you only look at a particular subsample,
  - be curious and critical.

  There are many modifications you may consider. The changes that you suggest should be motivated by some economic reasoning, they should not be completely artificial.
- Write down what you found: explain what did you try and motivate why. Discuss any potentially interesting findings.
- Excellent and insightful replications will be awarded +1% bonus.

Make sure to comment your code and make your best effort to adhere to some reasonable coding standards. Your code must be easy to read and it should take minimal effort to reproduce your results. Present your results in a coherent way and whenever possible make use of visualization.

[1] Please do not choose: Acemoglu, Daron, Simon Johnson, and James A. Robinson. 'The colonial origins of comparative development: An empirical investigation.' American economic review 91.5 (2001): 1369-1401. which is replicated in the `lecture4_2023.R` in the IS MUNI system. We will finish the Instrumental variables topic at the beginning of the next session.