ECON42720 Causal Inference and Policy Evaluation

Assignment 1: Selection on Observables

Ben Elsner

0.1 About this Assignment

This assignment teaches you how to apply the methods of Lectures 1-4 of the course (DAGs, regression, potential outcomes, matching).

- You can work on this assignments in groups of up to 4. I will not get involved in the group formation process.
- Please submit one assignment per group on Brightspace under *Assessment* > *Assignments*. Make sure that all the names of the group members are on the assignment.
- The assignment should be submitted as one pdf file. It should contain a write-up with your answers to the questions (including tables and figures) and the R code you used to generate the results. Do not use code chunks; instead, provide the entire code at the end of the document.
- Formatting: I recommend using Quarto, but it's not a must. I expect nicely formatted tables and figures (use ggplot2 or similar for figures, stargazer or similar for tables). Before AI became broadly available, students could get away with ugly tables and figures because producing nices ones takes time. Not any more.
- AI Policy. I expect that you work on the solutions yourself. Feel free to use AI to help with coding and editing, but I warn against using it to come up with solutions. Please include an AI statement that explains how you used AI in your assignment.
- I will provide feedback on Brightspace to the person who submitted the assignment. Please share the feedback and the grade with the group.

0.2 Tasks

0.2.1 DAGs

0.2.2 Simulation

0.2.3 Regression

Please load the dataset `jtrain3` from the `wooldridge` package (see Section 3 for how to do that). The dataset contains information on the training programs for workers in the US. The variable `re78` is the outcome of interest, the earnings in 1978. The variable `train` is a binary variable indicating whether the worker participated in the training program.

0.2.4 Matching

0.3 Some Hints

0.3.1 Loading a dataset

To load a dataset from the `wooldridge` package, use the following code:

```
install.packages("wooldridge") # install package if needed
library(wooldridge)
data("jtrain3")
df <- jtrain3 # do this because the data loads lazily
```

0.3.2 Running a logit regression

You can run a logit regression using the `glm` function. Here is an example:

```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
# Simulate data with 100 observations and 2 covariates
set.seed(123) # Set seed for reproducibility
n <- 100
X1 <- rnorm(n)
X2 <- rnorm(n)
D <- rbinom(n, 1, 0.5) # 50% are treated here
data <- data.frame(X1, X2, D)

# Run a logit regression

logit_model <- glm(D ~ X1 + X2, data = data, family = binomial(link = "logit"))
stargazer(logit_model, type = "latex", header=FALSE) # Print the results
```

Table 1

	Dependent variable:
	D
X1	−0.336
	(0.229)
X2	0.060
	(0.211)
Constant	−0.128
	(0.205)
Observations	100
Log Likelihood	−67.819
Akaike Inf. Crit.	141.639
Note:	*p<0.1; **p<0.05; ***p<0.01