

ECON42720 Causal Inference and Policy Evaluation

Feedback on Assignment 1

Academic Year 2023/24

Prof. Benjamin Elsner

1.1: DAGs and Potential Outcomes

Overall, the analysis here was very well done. There were some inaccuracies in the answers, or imprecise use of statistical or causal language. Please see some comments before. There were very few errors that were common; most points below address issues in the answers of one or two students only.

- 1a) Mediators and confounders need to have distinct boxes and arrows.
- 1a) Remember that the absence of an arrow also says something. Many of you drew distinct confounders and mediators, but often the confounders could plausibly affect the mediators. For example, well-being could be a mediator, but it is almost certainly affected by parents and peers, whose influences are typically confounders.
- There should be no mention of functional form when commenting on a DAG. This would only be relevant once we translate a DAG into a regression model or some other model.
- Some of you referred to multicollinearity in the DAGs. Multicollinearity is not a problem in a DAG; it only is a problem in a regression model and even there it is rarely a problem.
- 1b) it is advisable to write out the backdoor paths and then show which variables need to be controlled for.
- 1b) Careful with statements such as "we can adjust for the mediator if it is the only channel through which the treatment affects the outcome." If it is the (one and) only channel, you cannot control for it because it is perfectly correlated with the outcome. If there is a direct effect and exactly one indirect channel, then adjusting for the mediator would identify the direct effect.
- 1c) It is important to state what type of bias we have. Just stating that including or omitting a variable leads to an endogeneity problem is not enough (that's how people who were trained in the 90s and before approach causal inference, with not much success...). Tell us whether an inclusion or omission leads to confounding, collider or bad control bias or maybe some more sophisticated form of bias that doesn't have a name.
- 1c) Some of you identified a collider problem if educational attainment is downstream from health and is included as a control variable. This is not, strictly speaking, a collider unless it is directly affected by a confounder or the treatment. So it really depends on the DAG.
- 1c) Some of you stated in 1c that including or omitting a variable leads to an underestimation (or overestimation). In such instances you need to explain why this is the case. For this it's best to use the OVB formula (See lecture 2).
- 1d) When discussing the experimental design, it is mainly about randomisation, compliance, spillovers and attrition. No need to talk here about common support or the like.
- 1d) Randomisation at the individual level can work if there are no spillovers, but that is not very likely. Randomisation at the school level is better because spillovers are unlikely.

1.2 Empirical Analysis: Preparation and Data Inspection

Everyone got the analysis right, which is great. The detail in the answers varied across students. I do not believe that one needs to write several pages here, but it is important to explain in a few sentences what we see in a graph, why we see these differences and what these differences imply for the empirical analysis. It is important to learn to connect statistical explanations with interpretations that relate to the particular context. Later in life, when you communicate to policymakers or managers or whoever, they want to hear more than statistical arguments such as "X has a lower mean and a wider dispersion than Y". They want to know what this means in the context you study.

- b) and c): when commenting on common support and the distributions, it is welcome to provide intuition what it means not to have full common support and why the distributions are different. For example, the average family income of treated units is smaller because the government picked schools in poorer areas.
- d) It's important to understand why the propensity score distributions are so different. Treated and control units differ greatly in observable characteristics that predict the treatment status. In other words, among the untreated observations there are very few that look like treated observations; among the treated observations, there are very few that look like untreated observations.

1.3. Empirical Analysis:

Everyone did the analysis correctly and most provided excellent explanations and interpretations.

- b): When using the OVB formula – either the actual formula or its intuition in plain English – it is important to explain both components of the bias term, namely 1) how the omitted variable and the outcome are correlated, and 2) how the omitted variable and the regressor of interest are correlated. If both have the same sign, the bias is positive; if both have different signs, the bias is negative.
- d) There are many ways to perform balancing checks. Perhaps the most common ones in the literature are t-tests and a regression of the variable on a treatment indicator. But other tests were also ok.

1.4 Monte Carlo Simulation

Most of you got the code right. The answers varied in detail and accuracy. The goal here was to illustrate the bias-variance trade-off. The results show that the variance of the estimator is greater with $k=10$. In most cases it also produces an ATT that is closer to the true value. To see this, drawing the two distributions in the same graph would be helpful.

Coding feedback:

If you use Quarto or R Markdown, please suppress the warnings with the following code chunk:

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

If you use Quarto or R Markdown, it is useful to learn how to create latex tables from within R with stargazer.

- You need to install a tex distribution (the tinytex package in R does the trick)
- In the code chunk where the table appears, specify the option results='asis'
- Within the code chunk, when you use stargazer, specify the option type='latex'

```
model <- lm(wage ~ educ, data = wage1)
stargazer(model, header=FALSE, type='latex', title="Effect of Education on Wages")
```

Language feedback:

- Some refer to columns in a regression table as "models." I'm aware that some branches of statistics do that, but it's not strictly correct. There is only one correct model that generates the data in the population (and you don't know what that model looks). We estimate different regression specifications based on a random sample. In econometrics, we refer to columns in a regression table as "specifications" or simply "regressions."
- Verbs describing a causal relationship are "to have an effect on" and "to affect"; so A has an effect on B or A affects B; you cannot write A effects B.