Toulouse School of Economics

M1 INTERMEDIATE ECONOMETRICS

Causal inference

Koen Jochmans François Poinas

2025 — 2026

The identification problem

Let $D \in \{0, 1\}$ be a binary indicator that captures whether treatment is assigned or not.

The potential outcomes $Y(0)$ and $Y(1)$ are a pair of random variables that indicate the outcome of interest when the treatment was assigned or not.

The causal effect of the treatment is the difference

$$Y(1) - Y(0).$$

(In general, this is still a random variable).

The identification problem is to say something about causal effects from data on D and

$$Y = Y(1)\, D + Y(0)\, (1 - D)$$

alone. That is, we never observe both $Y(1)$ and $Y(0)$ together.

Comparison of means

We can compute

$$\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0).$$

This is the (population) slope coefficient in a regression of Y on D.

This equals

$$\begin{aligned}\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) &= \mathbb{E}(Y(1)|D=1) - \mathbb{E}(Y(0)|D=0) \\ &= \mathbb{E}(Y(1)|D=1) - \mathbb{E}(Y(0)|D=1) \\ &+ \mathbb{E}(Y(0)|D=1) - \mathbb{E}(Y(0)|D=0),\end{aligned}$$

that is,

$$\mathbb{E}(Y(1) - Y(0)|D=1) + \{\mathbb{E}(Y(0)|D=1) - \mathbb{E}(Y(0)|D=0).\}$$

In general,

$$\mathbb{E}(Y(0)|D=1) \neq \mathbb{E}(Y(0)|D=0),$$

so least-squares does not identify a causal effect without any further restrictions.

Selection into treatment

In observational data (economic) agents are making their own choices.

This leads to self-selection problems.

Consider a simple Roy model, where

$$D = \begin{cases} 1 & \text{if } Y(1) - Y(0) > 0 \\ 0 & \text{if } Y(1) - Y(0) \leq 0 \end{cases},$$

so that a unit selects into treatment when it pays off for him to do so.

Then

$$\begin{aligned} \mathbb{E}(Y(0)|D = 0) &= \mathbb{E}(Y(0)|Y(1) \leq Y(0)) \\ &\neq \mathbb{E}(Y(0)|Y(1) > Y(0)) = \mathbb{E}(Y(0)|D = 1) \end{aligned}$$

Randomization

Experimental data does not suffer from the selection problem.

Random assignment of treatment implies that

$$Y(0), Y(1) \perp\!\!\!\perp D,$$

so that potential outcomes are independent of treatment assignment.

Then

$$\begin{aligned}\mathbb{E}(Y|D=1) &= \mathbb{E}(Y(1)|D=1) = \mathbb{E}(Y(1)),\\ \mathbb{E}(Y|D=0) &= \mathbb{E}(Y(0)|D=0) = \mathbb{E}(Y(0)),\end{aligned}$$

and so

$$\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) = \mathbb{E}(Y(1) - Y(0))$$

identifies the average treatment effect.

Can again be estimated by simple regression.

Selection on observables

A weaker condition is a conditional-independence requirement given a set of control variables, X (confounding factors).

That is, selection into treatment can depend on X.

Then

$$Y(0), Y(1) \perp\!\!\!\perp D \,|\, X.$$

Now,

$$\begin{aligned}\mathbb{E}(Y|D=1,X) &= \mathbb{E}(Y(1)|D=1,X) = \mathbb{E}(Y(1)|X),\\ \mathbb{E}(Y|D=0,X) &= \mathbb{E}(Y(0)|D=0,X) = \mathbb{E}(Y(0)|X),\end{aligned}$$

and so

$$\mathbb{E}(Y(1)-Y(0)|X) = \mathbb{E}(Y|D=1,X) - \mathbb{E}(Y|D=0,X),$$

The two conditional expectations on the right-hand side here can be approximated by flexible linear regression. This motivates the use of multiple regression.

Noncompliance in experimental data

Experimental design where

$$Z = \begin{cases} 1 & \text{if treatment was assigned} \\ 0 & \text{if not} \end{cases},$$

and

$$D = \begin{cases} 1 & \text{if treatment was taken up} \\ 0 & \text{if not} \end{cases}.$$

So Z is randomly assigned but D may not be.

With non-compliance, $\mathbb{P}(Z \neq D) > 0$.

If the decision to deviate from assignment is related to the potential outcomes,

$$Y(0), Y(1) \perp\!\!\!\perp D,$$

will not hold.

From now on presume that

$$Y(1) - Y(0) =: \theta$$

is constant. Then

$$Y = Y(0) + (Y(1) - Y(0))\, D = Y(0) + \theta\, D.$$

Because Z is randomly assigned,

$$\begin{aligned}\mathbb{E}(Y|Z=1) &= \mathbb{E}(Y(0)) + \theta\, \mathbb{E}(D|Z=1)\\ \mathbb{E}(Y|Z=0) &= \mathbb{E}(Y(0)) + \theta\, \mathbb{E}(D|Z=0)\end{aligned}$$

and so

$$\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0) = \theta\, (\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)).$$

Therefore,

$$\theta = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)} = \frac{\text{cov}(Y,Z)}{\text{cov}(D,Z)}$$

provided that

$$\mathbb{E}(D|Z=1) \neq \mathbb{E}(D|Z=0).$$

This is an instrumental-variable solution.

Z is informative about D but affects Y only through its impact on D.