

Modern causal inference and treatment effect estimation using Stata

Eduardo Garcia Echeverri

APSA Annual Meeting and Exhibition, Vancouver, 2025

Causal inference

Objective:

Measure the **effect on an outcome** of **changing a treatment variable** under our control.

Examples:

- Does raising the **minimum wage** increase **unemployment**?
- Do **vaccines** reduce the **mortality** of a disease like COVID?
- Does **voter registration** lead to more political **participation**?
- Do **natural resources** worsen **ethnic violence**?

Challenge 1 – Counterfactuals are not observed

Outcomes if units had received different treatments are **unknown**:

- Baby's weight if the mother **had smoked more/less**.
- Unemployment rate if state **had not raised minimum wage**.
- Patient health outcomes if they **had not been vaccinated**.

⇒ **Formal framework (and assumptions)** for these counterfactuals:

- Rubin's **potential outcomes framework**.

Challenge 2 – Confounding factors

Randomized experiments are the golden standard.

- Oftentimes **infeasible or unethical**.

⇒ Rely on **observational data** to answer causal questions.

Problem:

Unaccounted variables that influence treatment and outcome.

- ↑ Mother's income ⇒ ↓ smoking and ↑ prenatal controls
- ↑ Insurance ⇒ ↑ prob. vaccine and ↑ health outcomes

⇒ Choose **controls** wisely – Imbens et al. (2015), Pearl (2018)

Potential outcomes framework – Rubin(1974)

Binary **treatment variable**:

$T_i = 1$: unit i **was treated**;

$T_i = 0$: unit i **was not treated**.

Potential outcomes:

$Y_i(1)$: **outcome** of unit i **if treated**;

$Y_i(0)$: **outcome** of unit i **if not treated**.

Remark: T_i , $Y_i(1)$, and $Y_i(0)$ are **random variables**.

Quantities of interest in causal analysis

Individual treatment effect:

$$ITE_i = Y_i(1) - Y_i(0).$$

Average treatment effect:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Average treatment effect on the treated:

$$ATT = \mathbb{E}[Y_i(1) - Y_i(0) | T_i = 1].$$

Conditional average treatment effect:

$$CATE = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x].$$

Fundamental problem of causal inference

Both potential outcomes (PO's) are never observed.

When unit i is treated:

- $Y_i(1)$ is observed, but $Y_i(0)$ is not.

When unit i is not treated:

- $Y_i(0)$ is observed, but $Y_i(1)$ is not.

Therefore, **we observe** $Y_i = (1 - T_i)Y_i(0) + T_iY_i(1)$

⇒ **We need assumptions** to estimate ATE, ATT, CATE ...

Stable unit treatment value assignment (SUTVA)

PO's for unit i **depend only on** T_i (and possibly X_i).

- **Don't depend on** PO's or treatment of **other units**.

Examples of SUTVA violation:

- **Spillover** effects:
 - Herd immunity in vaccines, marketplace effects, network effects, ...
- Changing treatment if I see peers getting treated.

Not testable in general.

Unconfoundedness (a.k.a selection on observables...)

Conditional on observables, treatment and PO's are **independent**:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i | X_i$$

Remark: This does not mean that $Y_i \perp\!\!\!\perp T_i | X_i$

Examples of violations:

- Selection on **unobservables** (u_i)
- Units that “benefit” more from treatment **select** more into it:

$$\text{Cor}(Y_i(1), T_i | X_i) > 0$$

Not testable in general.

Overlap

Units must have a **positive probability** of being treated and of not being treated.

$$0 < \Pr[T_i = 1|X_i] < 1$$

Remark: $\Pr[T_i = 1|X_i]$ is called the **propensity score**.

Assumption is **close to failing** \Rightarrow estimators become **unstable**.

Easily testable in Stata!

Causal inference in Stata

The models we'll study today require for **consistency**:

- SUTVA,
- unconfoundedness,
- and overlap.

Other parametric assumptions will be introduced when necessary.

Use **Stata** to obtain causal inference **parameters**, interpret results, and obtain **model diagnostics**.

Session 1: Cross-sectional (CS) data

teffects:

- Regression adjustment (RA)
- Inverse probability weighting (IPW)
- Augmented IPW (AIPW)
- IPW regression adjustment (IPWRA)
- Matching

telasso:

- Treatment effect estimation using LASSO
(*high-dimensional data*)

cate:

- Conditional average treatment effects (CATE)
(*treatment effect heterogeneity*)

Session 2: Repeated CS and panel data

`didregress` and `xtdidregress`:

- Difference-in-differences (DID) models
- Difference-in-difference-differences (DDD) models

`hdidregress` and `xthdidregress`:

- Heterogeneous DID models
(*treatment effect heterogeneity*)

References

1. Heckman, J. J., and R. Pinto. 2022. Causality and econometrics. Working Paper, National Bureau of Economic Research.
2. Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*.
3. Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*.
4. Pearl, J., and D. MacKenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.