# VU Empirical Economics: Impact Evaluation

## Causal inference and counterfactuals


Andreas Steinmayr

University of Innsbruck


Fall 2023

*Core reading: Gertler, chapters 3; Cunningham, chapters 3 and 4*

# Causal inference

*Remember:* Impact evaluations seek to answer cause-and-effect questions precisely.

The **causal impact** of a program on an outcome is the difference between

- the outcome with the program
- and the same outcome without the program

To put it another way, we would like to measure the outcome **at the same point in time for the same unit**, but in **two different states of the world**.

Obtaining these measures is impossible because we only observe one state of the world. We need a framework to think about the problem.

# RCM: Observed and potential outcomes

- In *traditional* econometrics, we use **observed outcomes** $Y$
- The literature on program evaluation builds on the notion that each member of the population can be characterized by (two) **potential outcomes**:

$$\begin{aligned} Y^0 &: \text{ potential outcome if not treated} \\ Y^1 &: \text{ potential outcome if treated} \end{aligned}$$

- Causality is based on **comparison of potential outcomes**
  - Requires counterfactual thinking
- Statistical framework: *Rubin Causal Model* (Rubin 1974)

# RCM: Observed and potential outcomes

- The treatment indicator is defined as

$$W = \begin{cases} 0 & \text{if not treated} \\ 1 & \text{if treated} \end{cases}$$

- We will stick to binary treatments for most of the course but at times also consider situations in which the treatment variable can take different values ("treatment intensities").

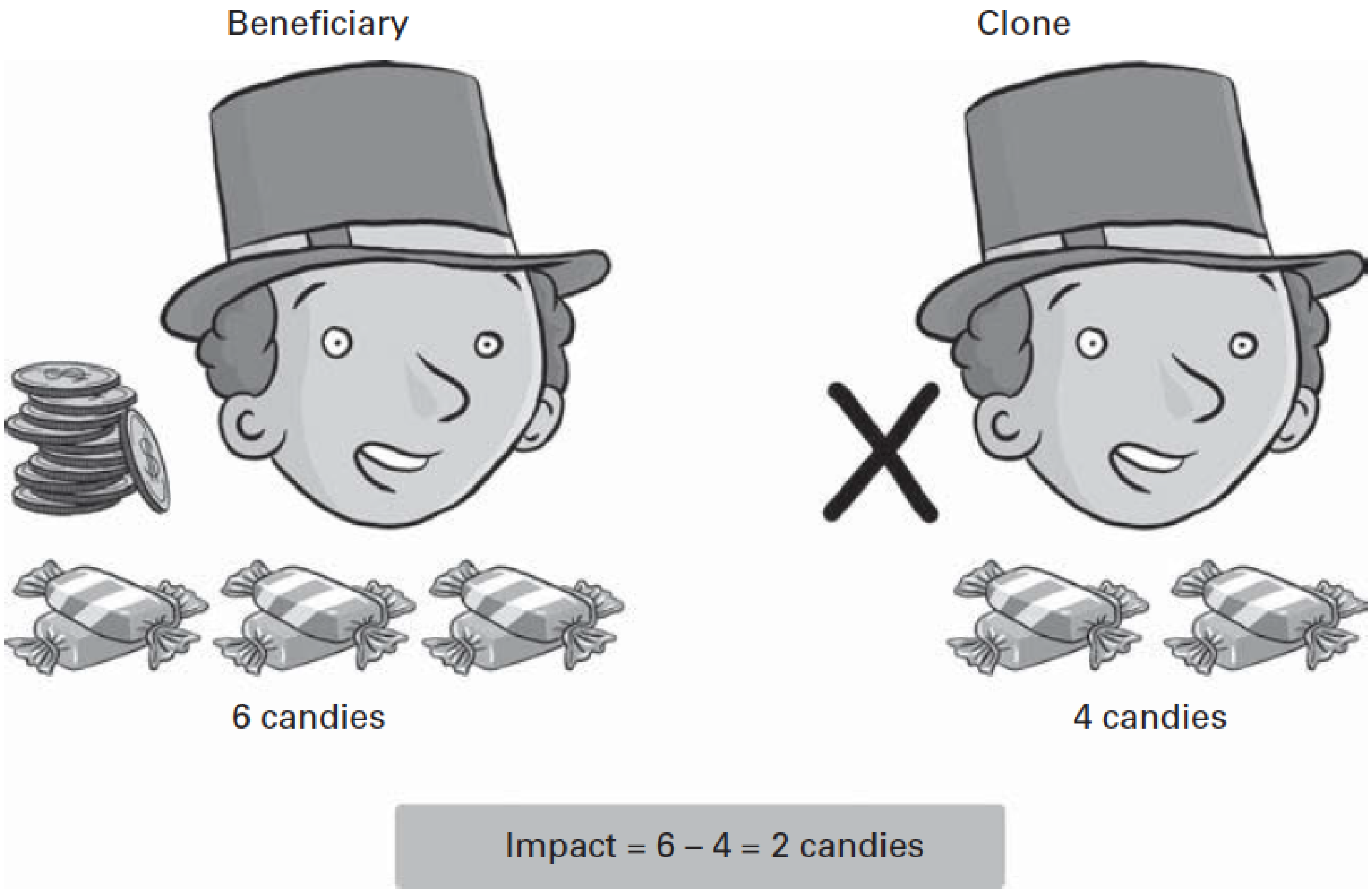
- After treatment is applied, only one of the potential outcomes is realized:

$$Y = Y^W = \begin{cases} Y^0 & \text{if W=0} \\ Y^1 & \text{if W=1} \end{cases}$$

# Stable Unit Treatment Value Assumption (SUTVA)

- Treatment of some unit (e. g., a person) only affects that persons outcome – this is often referred to as SUTVA.
- No multiple versions of the program (no good/bad pills)
- Rules out general equilibrium effects - program evaluation usually is about partial equilibrium analysis.
  - For example, the fact a some training raises one person's skill level might imply that another person looses her job if the number of jobs is fixed.

# The perfect clone



# RCM: Causal effect

- The **treatment effect** is the difference of the two hypothetical outcomes:

$$Y^1 - Y^0$$

- While we can think hypothetically about this quantity, **we can't observe it for any specific member of the population**.
- This is the fundamental evaluation problem
- The unobserved outcome is what we call the $Missing\ Counterfactual$
- The evaluation problem is a missing data problem

# Manipulation

Causal questions require careful description of manipulation to be well defined:

"She did not get this job because she is a woman."

What is the manipulation?

- Change chromosome after conception
- Change gender on written job application

Very different manipulations, and probably different causal effects!

Discussion: If we cannot think of a manipulation, why are we interested in the effect?

# RCM: Average Treatment Effect

- Instead of individual treatment effects, we can focus on the average effect of the treatment.
- We call this the **average treatment effect** (ATE)

$$\tau_{ATE} = \mathrm{E}[Y^1 - Y^0]\,,$$

- Note that we make no assumption that the effect is the same for everyone.
- We have to think about for which group we are identifying the ATE.

# RCM: Average Treatment Effect

- The average treatment effect, $\tau_{ATE}$ is an unknown population parameter. Thus, we assign a Greek letter to it, as in the case of a population mean where we define $\mathrm{E}(y) = \mu$. The evaluation literature prefers using acronyms (as in: "We estimate the ATE.")
- There is a crucial difference between the population mean, $\mathrm{E}(y)$, and the average treatment effect, $\mathrm{E}[Y^1 - Y^0]$:
  - With a random sample from the population, $\mu$ can be estimated by the sample analog of $\mathrm{E}(y)$: $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$.
  - $\tau_{ATE}$ cannot be estimated by a sample average without additional assumptions. (Why?)

# RCM: Average Treatment Effect on the Treated

- Another population quantity of interest is the **average treatment effect on the treated** (ATT):

$$\tau_{ATT} = \mathrm{E}(Y^1 - Y^0 | W = 1)\,,$$

  which is the expected effect of treatment on the outcome for a randomly drawn member of the sub-population that received the treatment.

- The ATT can't be estimated by a sample average without additional assumptions either.

# RCM: Average Treatment Effect on the Non-treated

- Another population quantity of interest is the **average treatment effect on the non-treated** (ATNT):

$$\tau_{ATNT} = \mathrm{E}[Y^1 - Y^0 | W = 0] \,,$$

which is the expected effect of treatment on the outcome for a randomly drawn member of the sub-population that did not receive the treatment.

# Exercise

Use the following dataset to estimate the ATE, ATT, and ATNT. Note: Quantities in italics are not observed in real world data but you can use them here.

| $i$ | $W_i$ | $Y_i^0$ | $Y_i^1$ | $\Delta_i = Y_i^1 - Y_i^0$ |
|---|---|---|---|---|
| 1 | 0 | 6 | *8* | |
| 2 | 0 | 6 | *9* | |
| 3 | 0 | 9 | *11* | |
| 4 | 1 | *8* | 12 | |
| 5 | 1 | *7* | 10 | |
| 6 | 1 | *2* | 6 | |
| 7 | 1 | *10* | 13 | |
| 8 | 1 | *5* | 9 | |
| 9 | 0 | 7 | *10* | |
| 10 | 0 | 10 | *12* | |

# No information bounds

- What can we say about causal parameters without data and assumption?

- Even without assumptions and data, we know that the treatment effects are in some interval if the expectations of the outcome variable of interest are bounded

- Assumption BS: Bounded support of potential outcomes: $\mathrm{E}[Y^0|W = 1] \in [\underline{Y}, \overline{Y}]$ and $\mathrm{E}[Y^1|W = 0] \in [\underline{Y}, \overline{Y}]$

# No information bounds: example

- Example: ATT of college degree on labor supply. Labor supply is a binary variable and thus bounded.

# No information bounds: example

- Example: ATT of college degree on labor supply. Labor supply is a binary variable and thus bounded.

- Without any data we can say

$$\tau_{ATT} = \mathrm{E}[Y^1|W=1] - \mathrm{E}[Y^0|W=1]$$

- Upper bound of effect: $\overline{\tau_{ATT}} = \overline{Y} - \underline{Y}$

- Lower bound of effect: $\underline{\tau_{ATT}} = \underline{Y} - \overline{Y}$

- Length of bounds is $2 * (\overline{Y} - \underline{Y})$

# No information bounds: example

With observable data containing $Y$ and $W$ the bounds for the ATT shrink to:

# No information bounds: example

With observable data containing $Y$ and $W$ the bounds for the ATT shrink to:

- Upper bound of effect: $\overline{\tau_{ATT}} = \mathrm{E}[Y|W = 1] - \underline{Y}$
- Lower bound of effect: $\underline{\tau_{ATT}} = \mathrm{E}[Y|W = 1] - \overline{Y}$
- Length of bounds is $(\overline{Y} - \underline{Y})$

# No information bounds: example

With observable data containing $Y$ and $W$ the bounds for the ATT shrink to:

- Upper bound of effect: $\overline{\tau_{ATT}} = \mathrm{E}[Y|W=1] - \underline{Y}$
- Lower bound of effect: $\underline{\tau_{ATT}} = \mathrm{E}[Y|W=1] - \overline{Y}$
- Length of bounds is $(\overline{Y} - \underline{Y})$

Data reduce uncertainty (interval) by 50%
Assumptions have to reduce the other 50%

# The naïve estimator

- The estimation problem for both $\tau_{ATE}$ and $\tau_{ATT}$ results from the fact that we never observe both $Y^0$ and $Y^1$ at the same time.
- We only observe realized outcomes and the treatment indicator:

$$\begin{aligned} y &= (1-w)\,y_0 + w\,y_1 \\ &= y_0 + w(y_1 - y_0) \end{aligned}$$

- With a sample of $(Y, W)$, why can't we take the simple difference

$$\hat{\tau}_N = \bar{y}_{w=1} - \bar{y}_{w=0}$$

- which would correspond to the following regression:

$$y = \beta_0 + \beta_1 w + u$$

# Example: Roy model

Suppose there are only two occupations in the world: economists and accountants. Suppose non-wage aspects of the jobs are the same.

Earnings of accountants are given by:

$$Y^0 \sim N(60000, 5000)$$

Earnings of economists are:

$$Y^1 \sim N(60000, 10000)$$

In addition, we assume the correlation between accounting and economist wages is high: 0.84. If you're going to be a good economist, it's very likely you'll also be good at accounting.

# Example: Roy model

Now let's build a model of occupation selection. Because non-wage aspects of both jobs are the same, the worker picks the one that pays the most. Her observed earnings are

$$y_i = max(y_i^0, y_i^1)$$

Note that here $Y_i$ is written in lower case because it is the realization of a random variable. Because both $Y_i^0$ and $Y_i^1$ are lower-case on the right-hand side of the equation, we are assuming that our agent knows what her earnings would be in each occupation before she decides to be an economist or accountant. Let $W_i = 1$ indicate she chooses economics. Now, since we are devising the model, we can do something we can't do in real life: observe the potential earnings of the entire population.

# Example: Roy model

| | Accountants | Economists | Total |
|---|---|---|---|
| Accounting earnings | 58,346 | *61,648* | *59,996* |
| Economics earnings | *54,469* | 65,519 | *59,991* |
| N | 500,245 | 499,755 | 1,000,000 |

Quantities in italics are the unobserved counterfactuals and not observed in real world data.

We, as economists, want to know what impact choosing to be an economist has on economists' earnings relative to the counterfactual of them being accountants. We have defined treatment as being an economist, so what treatment parameters are we interested in?

# Example: Roy model

The naïve estimator gives us

$$\begin{aligned}\hat{\tau}_N &= \bar{y}_{w=1} - \bar{y}_{w=0} = \bar{y}^1_{w=1} - \bar{y}^0_{w=0} \\ &= 65,519 - 58,346 = 7,173\end{aligned}$$

This is naïve as it assumes that a good estimate of economists' counterfactual accounting earnings is the observed accountants' accounting earnings.

Since we got to observe all potential outcomes in this exercise, we can calculate what the real gain from becoming economists is for the people who become economists.

$$\hat{\tau}_{ATT} = \bar{y}^1_{w=1} - \bar{y}^0_{w=1} = 65,519 - 61,648 = 3,871$$

# Example: Roy model

Alternatively, you could have estimated the impact of becoming economists for those who actually became accountants

$$\hat{\tau}_{ATNT} \quad = \quad \bar{y}^1_{w=0} - \bar{y}^0_{w=0} = 54,469 - 58,346 = -3,877$$

It looks like those who became accountants made the right choice: they would have been worse off if they had become economists! This again shows why the naïve estimator is so bad.

# Example: Roy model

Finally, you may want to know what the impact would be if we made everyone become economists.

$$\hat{\tau}_{ATE} = \bar{y}^1 - \bar{y}^0 = 59,991 - 59,996 = -5$$

All three of these parameters are meaningful, and our naïve estimator gives us none of them.

# The selection problem

- This naïve estimator implicitly assumes that

$$\begin{aligned} \mathrm{E}[Y|W=1] &= \mathrm{E}[Y^1|W=1] = \mathrm{E}[Y^1] \\ \mathrm{E}[Y|W=0] &= \mathrm{E}[Y^0|W=0] = \mathrm{E}[Y^0] \end{aligned}$$

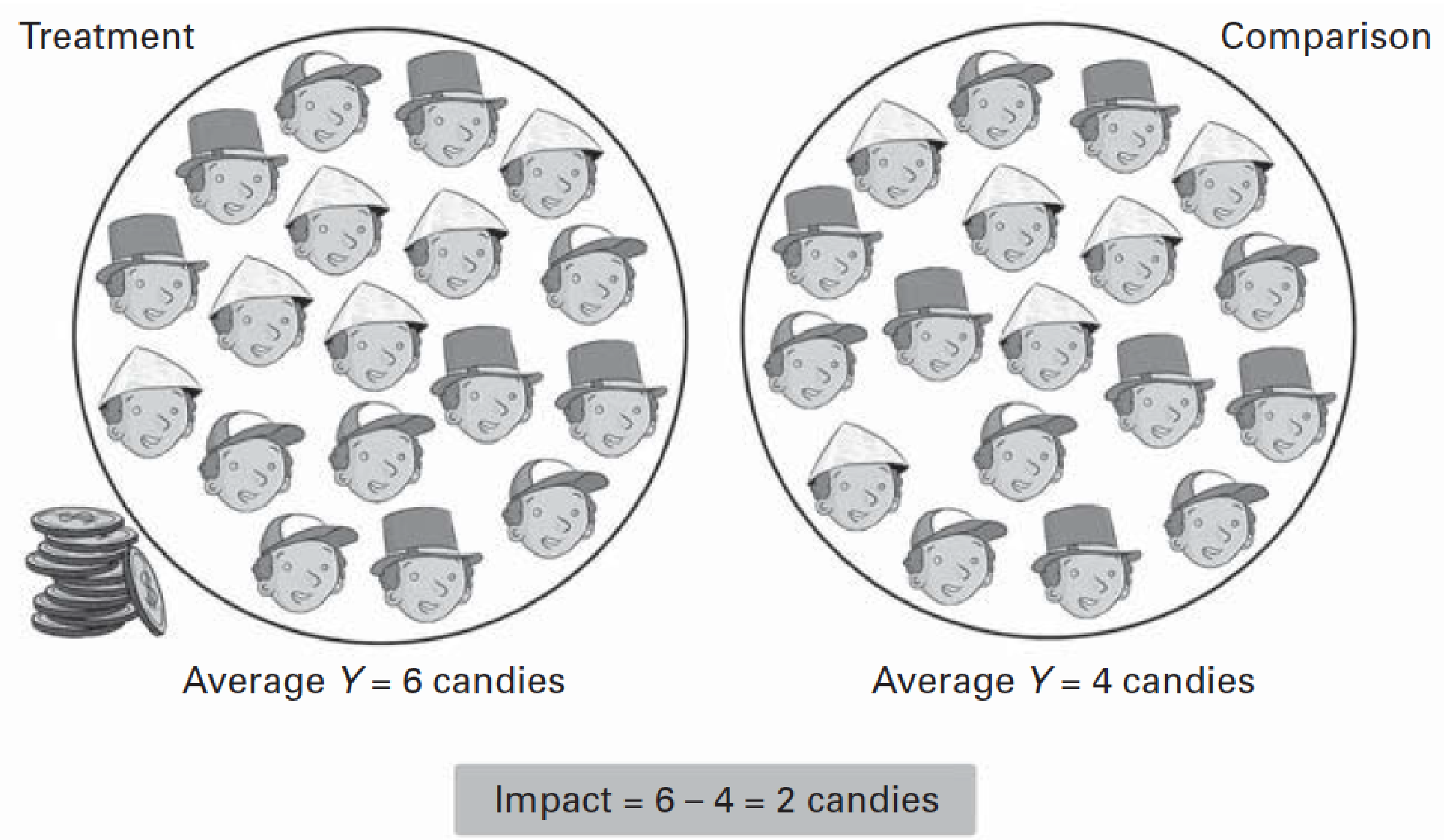
- In traditional econometrics, we called this *ceteris paribus*

- We call violations of these assumptions the **Selection Problem**

# Exercise

- Use the naïve estimator for the small dataset.
- Study the potential outcomes. Is the assumption behind the naïve estimator fulfilled?

# The selection problem: a valid comparison group



# The selection problem

- Dealing with the selection problem is often the biggest challenge in policy evaluation
- The selection problem arises because we only observe certain/selected people in the treated state
  - Participants are a nonrandom sample from the eligible population
  - They are different from those not treated
  - They are also different from themselves prior to the start of the treatment
- *Different* refers to: they have different potential outcomes
- **Why did these people get treated while others did not?**

# Two types of selection

- **Selection on Observables**
  - Participants are different from non-participants in terms of observable characteristics, i.e. characteristics for which we have measures in our data.

- **Selection on Unobservables**
  - Participants are different from non-participants in terms of unobservable characteristics.

- The type of selection has important implications for the empirical methods we can use.

- How did we think about selection in our $standard$ econometrics class?

# No information bounds

- What can we say about causal parameters without data and assumption?
- Manski (1989, 1990) and Robins (1989) show that even without assumptions and data, we know that the treatment effects are in some interval if the expectations of the outcome variable of interest are bounded
- Assumption BS: Bounded support of potential outcomes: $\mathrm{E}[Y^0|W=1] \in [\underline{Y}, \overline{Y}]$ and $\mathrm{E}[Y^1|W=0] \in [\underline{Y}, \overline{Y}]$

# No information bounds: example

- Example: ATT of college degree on labor supply. Labor supply is a binary variable and thus bounded.

- Without any data we can say

$$\tau_{ATT} = \mathrm{E}[Y^1|W=1] - \mathrm{E}[Y^0|W=1]$$

- Upper bound of effect: $\overline{\tau_{ATT}} = \overline{Y} - \underline{Y}$

- Lower bound of effect: $\underline{\tau_{ATT}} = \underline{Y} - \overline{Y}$

- Length of bounds is $2 * (\overline{Y} - \underline{Y})$

# No information bounds: example

- With observable data containing $Y$ and $W$ the bounds for the ATT shrink to:
  - Upper bound of effect: $\overline{\tau_{ATT}} = \mathrm{E}[Y|W=1] - \underline{Y}$
  - Lower bound of effect: $\underline{\tau_{ATT}} = \mathrm{E}[Y|W=1] - \overline{Y}$
  - Length of bounds is $(\overline{Y} - \underline{Y})$

# No information bounds: example

- With observable data containing $Y$ and $W$ the bounds for the ATT shrink to:
  - Upper bound of effect: $\overline{\tau_{ATT}} = \mathrm{E}[Y|W = 1] - \underline{Y}$
  - Lower bound of effect: $\underline{\tau_{ATT}} = \mathrm{E}[Y|W = 1] - \overline{Y}$
  - Length of bounds is $(\overline{Y} - \underline{Y})$
- Data reduce uncertainty (interval) by 50%
- Assumptions have to reduce the other 50%

# Literature


- Manski, C. F. (1989): The Anatomy of the Selection Problem, *Journal of Human Resources*, 24, 343-360.
- Manski, C. F. (1990): Nonparametric Bounds on Treatment Effects, *American Economic Review, Papers and Proceedings*, 80, 319-323.
- Robins, J. M. (1989): The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies", In: Sechrest, L., H. Freeman, A. Mulley (eds.), *Health Service Research Methodology: A Focus on Aids*, 113-159, Washington, D.C.: Public Health Service, National Center for Health Services Research.
- Roy, A. D. (1951): Some Thoughts on the Distribution of Earnings, *Oxford Economics Papers*, 3(2), 135-146.
- Rubin, D.B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, 66, 688-701.