Selection Bias and how to fight it

Eduard Bukin

Slides are based on two recommended readings. Read them if you can!

1. Ch. 1 in (Angrist & Pischke, 2014)
2. Ch. 2 in (Angrist & Pischke, 2009)

What do econometricians do?

- Econometricians search and reveal the **causal effects**
- Often refereed as **Average Treatment Effect (ATE)**:

$$\rho$$

- But the ATE is always hidden from us behind the **Selection Bias**!

Rubin’s causal model

(Holland, 1986; Rubin, 1974, 1977)

Potential outcomes framework

$D_i = \{0, 1\}$ is a treatment that causes a change in the **actual outcome** Y_i;

Y_{0i} and Y_{1i} are two potential outcomes for an individual i;

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Potential outcome is **what we would measure if we could go back in time and change a person's treatment status.**

We observe:

- $Y_i = Y_{1i}$, when $D_i = 1$
- $Y_i = Y_{0i}$, when $D_i = 0$

Causal effect of a treatment

- Is the difference between two potential outcomes:

$$\rho = Y_{1i} - Y_{0i}$$

Depending on what we observe as a factual:

- Y_{0i} is the counterfactual for Y_{1i}
- Y_{1i} is the counterfactual for Y_{0i}

Actual outcome

$$
\begin{aligned}
Y_i &= \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \\
&= Y_{0i} + \rho \cdot D_i
\end{aligned}
$$

Actual outcome: single path that an individual walks.

Conditional Expectation:

Actual outcome is Conditional on treatment

- We use "$|$" to denote **conditional** on something (D_i);
- $[Y_i|D_i]$ means actual outcome Y_i conditional on D_i
- When $D_i = 0, \quad [Y_i|D_i = 0] = Y_{0i} + 0$
- When $D_i = 1, \quad [Y_i|D_i = 1] = Y_{0i} + (Y_{1i} - Y_{0i})$

Comparing of two individuals

We have two individuals with actually observed outcomes:

- $[Y_i|D_i = 1]$
- $[Y_i|D_i = 0]$

$$\underbrace{[Y_i|D_i = 1] - [Y_i|D_i = 0]}_{\text{Observed difference}}$$

$$\underbrace{[Y_i|D_i = 1] - [Y_{0i}|D_i = 1]}_{\text{Average treatment effect on treated}}$$

$$+ \underbrace{[Y_{0i}|D_i = 1] - [Y_i|D_i = 0]}_{\text{Selection bias}}$$

$$\underbrace{[Y_i|D_i = 1] - [Y_i|D_i = 0]}_{\text{Observed difference}} = \underbrace{[Y_i|D_i = 1] - [Y_{0i}|D_i = 1]}_{\text{Average treatment effect on treated}} + \underbrace{[Y_{0i}|D_i = 1] - [Y_i|D_i = 0]}_{\text{Selection bias}}$$

Differences between treated and not treated are always affected by the Selection Bias

The origin of the selection bias

In:

$$[Y_i|D_i=1]-[Y_i|D_i=0]=[Y_i|D_i=1]-[Y_{0i}|D_i=1] \\ +[Y_{0i}|D_i=1]-[Y_i|D_i=0]$$

- **actual** outcome of NO-treatment in NOT treated is the same as **potential** outcome of NO-treatment in NOT treated.

$$[Y_i|D_i=0]=[Y_{0i}|D_i=0]$$

However:

- **potential** outcome of NO-treatment in **treated** is **NOT** the same as **potential** outcome of NO-treatment in NOT treated.

$$[Y_{0i}|D_i=1]\neq[Y_{0i}|D_i=0]$$

Selection bias in the nutshell

- Selection bias arises from the lack of compatibility:
  - when we compare phenomena that are not comparable;
  - like **apples** and **oranges**;
- As every person is unique, the potential outcomes of treatment and no treatment are different between people.
- Selection bias arises when we do not have the Ceteris Paribus!

Example 1

Does conditional cash transfers (CCT) cause a reduction in children wasting (Z score)?

Two households

Household ($i = 1$):

1. Received CCT (treatment):
   - $D_i = 1$
2. Observed wasting: **2 SD**;
   - Actual outcome = Potential outcome when **treated**;
   - $$[Y_{i=1}|D_{i=1} = 1] = [Y_{1,i=1}|D_{i=1} = 1] = 2$$

Household ($i = 2$):

1. No CCT (no treatment):
   - $D_i = 0$
2. Wasting: **1 SD**;
   - Actual outcome = Potential outcome when **NOT treated**;
   - $$[Y_{i=2}|D_{i=2} = 0] = [Y_{0,i=2}|D_{i=2} = 0] = 1$$

What are the similarities between two household?

What are the differences?

Comparing two households (the difference)

$$\begin{aligned}\text{Difference between two HH} &= \\ &= [Y_{i=1}|D_{i=1}=1] - [Y_{i=2}|D_{i=2}=0] \\ &= 2 - 1 \\ &= 1\end{aligned}$$

- Is this an Average Treatment Effect (ATE)?
- Is this the Treatment Effect on Treated?
- Vote?

Potential Outcomes of two households (1/2)

		Household $i = 1$ (Treated)	Household $i = 2$ (Not treated)
Potential outcome without CCT:	Y_{0i}	-2 Not observed	1 Observed
Potential outcome with CCT:	Y_{1i}	2 Observed	1 Not observed
Actual treatment status:	D_i	1 Observed	0 Observed
Actual outcome:	Y_i	2 Observed	1 Observed
Treatment effect on treated:	$Y_{1i} - Y_{0i}$	$2 - (-2) = 4$ Not observed	$1 - 1 = 0$ Not observed

What can we actually observe? What (from above) can we NOT observe in the real world?

Potential outcomes are different for two HHs.

- Why are they different that so?

Potential Outcomes of two households (2/2)

		Household $i = 1$ (Treated)	Household $i = 2$ (Not treated)
Potential outcome without CCT:	Y_{0i}	-2	1
Potential outcome with CCT:	Y_{1i}	2	1
Actual treatment status:	D_i	1	0
Actual outcome:	Y_i	2	1
Effect of treatment on treated:	$Y_{1i} - Y_{0i}$	$2 - (-2) = 4$	$1 - 1 = 0$

Treatment causes different effects:

1. Effect of treatment on the treated:

$$\text{ETT}_i = Y_{1i} - Y_{0i}$$
$$\text{ETT}_1 = Y_{1,i=1} - Y_{0,i=1} = 2 - (-2) = 4$$
$$\text{ETT}_2 = Y_{1,i=2} - Y_{0,i=2} = 1 - 1 = 0$$

2. Average Treatment Effect (ATE):

$$\text{ATE} = E[\text{ETT}_i] = E[Y_{1i} - Y_{0i}] = \frac{1}{2}[2 - (-2) + 1 - 1] = 2$$

But the difference in the actual outcomes shows a biased effect: $Y_{1,i=1} - Y_{0,i=0} = 2 - 1 = 0$

Comparing two households (the bias)

$$\begin{aligned}\text{Diff. between two HH} &= [Y_{i=1}|D_{i=1}=1] - [Y_{i=2}|D_{i=2}=0] \\ &= \underbrace{[Y_{i=1}|D_{i=1}=1] - [Y_{0,i=1}|D_{i=1}=1]}_{\text{Effect of treatment on treated (ETT)}} \\ &\quad + \underbrace{[Y_{0,i=1}|D_{i=1}=1] - [Y_{i=2}|D_{i=2}=0]}_{\text{Selection bias}}\end{aligned}$$

$$\text{Diff. between two HH} = \underbrace{2-(-2)}_{4} + \underbrace{(-2)-1}_{\text{-3}} = \underbrace{1}_{\text{Biased effect}}$$

- True causal effect of treatment on treated in $i=1$ is 4 and in $i=2$ is 0;
- But comparing two groups does not reveal this!
- Instead, we have a negative bias of a HH who is OK without treatment.
- Such negative selection bias can mask true causal effect completely.

Example 2

Does conditional cash transfers (CCT) cause **on average** a reduction in children wasting (Z score)?

- Let us compare **two groups of households**;
- One group is from one village that received CCT; another group is from another village that did not receive CCT.
- What could be the similarities between the two groups?
- What could be the difference between the two groups?
- If groups averages are different, does it mean that CCT caused this difference?

Average treatment effect

The constant-effects assumption!

$$Y_{1i} = \rho + Y_{0i}$$

- The treatment has a constant effect ρ on all individuals.

When we reveal the causal effect ρ we assume that it is constant for all treated an not treated individuals.

Bias of the group means difference

$$\text{Diff. in group means} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

$$= \underbrace{E[Y_i|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average treatment effect (ATE)}}$$

$$+ \underbrace{E[Y_{0i}|D_i = 1] - E[Y_i|D_i = 0]}_{\text{Selection bias}}$$

$$\text{Diff. in group means} = \rho + \text{Selection bias}$$

Because:

- *average outcome* of no-treatment in **not treated**is the same as *average potential outcome* of **no-treatment in not treated**.

$$E[Y_i|D_i = 0] = E[Y_{0i}|D_i = 0]$$

However:

- *average potential outcome* of no-treatment in **treated** is **NOT** the same as *average potential outcome* of **no-treatment in not treated**.

$$E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$$

Selection Bias: conclusions

- Mortal enemy of the causal inference: leads to interpreting naive difference as causal effects.
- **Exists in any comparison, where there are systematic difference between the groups compared.**
- Appears when individuals are being “selected” for the comparison, thus:
  - on average, individuals are not the same, or
  - there is no Ceteris Paribus.

Example 3: (Wakefield et al., 1998)

What is wrong with (Wakefield et al., 1998)?

Remember: (1) treatment is the MMR vaccination. (2) outcome is the autism/inflation.

Treated group:

- 12 children with bad symptoms (all are vaccinated);

Counter factual:

- 12 “random” children with same age and gender (not vaccinated);

Comparison:

- Mean prevalence of autism and inflammations by MMR vaccination status;

- What is wrong?
  - Ideas?…
  - Counterfactuals are not the same as treated.
  - Counterfactuals do not represent the population (nearly everyong is vaccinated against MMR).
  - Selection bias affects the means comparison.
  - Does not reviel the causal effect.

Solutions to the Seletion Bias

> Mathematically, there is no solution to the selection bias.

1. We cannot rearrange numbers or variables to resolve it.

2. Once a individual walks the path, it cannot go back and take another turn.

> KWAI CHANG CAINE: What happens in a man’s life is already written. A man must move through life as his destiny wills.

> OLD MAN: Yet each is free to live as he chooses. Though they seem opposite, both are true.

> Kung Fu, Pilot (Angrist & Pischke, 2014)

Solutions to the Selection Bias

The only way to resolve it is:

- To **design the research** so that **the design eliminates the selection bias**.

Using **econometrics** enhanced by **statistics** and appropriate **research design**, we can:

- ensure the Ceteris Paribus by
- “making” groups of comparison as similar as possible
- and controlling the differences.

Furious Five econometric methods

For ensuring ceteris paribus econometricians use:

- Random assignment (RCT)
- Regression
- Instrumental Variable
- Difference-in-difference
- Regression Discontinuity Design

The power of a random assignment

Selection bias is the bias that appear when individuals are **selected** to treatment and control groups.

Random assignment kills “selection” by randomly assigning groups of treatment and control.

Random assignment makes treatment D_i independent of potential outcomes.

Random assignment (1/3)

Remember, we **had** to non randomly assigned groups:

$$\text{Diff. in group means} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$
$$= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average causal effect}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}},$$

$$\text{where } E[Y_i|D_i = 0] = E[Y_{0i}|D_i = 0], \text{ but}$$
$$E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$$

The random assignment of D_i makes:

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$$

Random assignment (2/3)

Now with $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$, we have:

$$\text{Diff. in group means} = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\rho}$$

$$+ \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{0}$$

$$= \underbrace{E[Y_{0i} + \rho|D_i = 1]}_{E[Y_{1i}|D_i=1]} - E[Y_{0i}|D_i = 0]$$

$$= \rho + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{0}$$

$$= \rho = \text{ATE}$$

Random assignment (3/3)

With random assignment,

- $\text{Diff. in group means} = \text{ATE}$ works because of
- **the Law of Large Numbers (LLN)**.

1. Provided that two samples are large enough for the LLN to work

2. LLN ensures that such large random samples asymptotically approximate the population.

3. Thus, samples are also same between each other.

4. **Research Design + LLN**, ensures that $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$.

The Randomized Control Trials (RCT)

Randomized control trials is a powerful tool of causal inference.

- Consider learning more about it!

Key redings on RCT

Key papers:

1. 1. Using Randomization in Development Economics Research: A Toolkit
2. 2. The Econometrics of Randomized Experiments

RCT in development:

1. 3. Conditional Cash Transfers : Reducing Present and Future Poverty (book)
2. 4. Causal Inference in Statistics, Social, and Biomedical Sciences An Introduction (book)

RCT in economics:

1. See reference in (Angrist & Pischke, 2009, Chapter 2; 2014, Chapter 1; Athey & Imbens, 2017)

RCT’s criticism

This research design is not ultimate and one needs to be critical to it as well!

See (Deaton & Cartwright, 2018):

- RCT does not ultimately equalize everything because of the sample size and randomization strategy. Other factors (covariates) must be controlled for.
- External validity of the RCT could be very limited.
- Building RCT should include prior knowledge.
- RCT’s finding may be contemporary.

Food for thought (Homework)

Homework

Watch these videos on youtube and read

Video 1: Selection Bias or this link: https://youtu.be/6YrIDhaUQOE

Video 2: Randomized Trials or this link: https://youtu.be/eGRd8jBdNYg

Read:

(Angrist & Pischke, 2014, Chapter 1; optional Angrist & Pischke, 2009, Chapter 2)

Finish the in-class exercise

Homwork: Discuss the following causal questions

1. Many farms, particularly in Europe, are small and owned and run by families. In Eastern Europe, Soviet legacy left large scale farm. Are small scale farms more efficient and productive then the large one?
2. What is the effect of Conditional Cash Transfers rather then support with goods in kind on the extreme poverty in developing countries?
3. What is the effect of Global Food prices surge on the number of the food security in the low income countries?

For each of these questions answer the following:

- What is the outcome variable and what is the treatment?
- Define the counterfactual outcomes Y_{0i} and Y_{1i} .
- What plausible causal channel(s) runs directly from the treatment to the outcome?
- What are possible sources of selection bias in the raw comparison of outcomes by treatment status?
- Does the selection bias overestimate the difference or underestimates it?

Takeaways / Exam topics that require understanding

Takeaways:

1. Average treatment effect (ATE) and Effect of Treatment on Treated (ETT);
2. Selection bias of means comparison and the lack of Ceteris Paribus;
3. What is the Ceteris Paribus (Watch a video is needed);
4. Actual and potential outcomes framework;
5. Factual and Counterfactual;
6. Role of research design in fighting with the selection bias;
7. **Furious Five** econometric methods;
8. Random assignment;

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press. http://doi.org/10.1515/9781400829828

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of field experiments* (pp. 73–140). Elsevier. http://doi.org/10.1016/bs.hefe.2016.10.003

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. http://doi.org/10.1016/j.socscimed.2017.12.005

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. http://doi.org/10.1080/01621459.1986.10478354

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. http://doi.org/10.1037/h0037350

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, *2*(1), 1–26. http://doi.org/10.3102/10769986002001001

Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., Malik, M., … Walker-Smith, J. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637–641. http://doi.org/10.1016/s0140-6736(97)11096-0