Notes on Lab Session 3

Giulia Tani
https://tanigiulia.github.io/

TSE - MSc course in Program Evaluation
January 2026

The key issue of observability

- For observation i, we **never** observe the impact of treatment $D \in \{0, 1\}$ on outcome Y:

$$Y_i(1) - Y_i(0)$$

- What we observe is:

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = \underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1)}_{\text{ATT}}$$
$$+ \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{Selection bias}}$$

- **Selection bias**: treated and untreated units would have had different outcomes *even without treatment*
- Example: what is the causal effect of installing speed cameras on crash rates?
  - » What we see: roads with cameras have 12 crashes/year, roads without cameras have 6.
  - » Naive conclusion: cameras double crashes.
  - » Selection bias: cameras are installed on risky roads, where crashes were already high (dangerous intersections, speeding corridors, ...).

Example

- Problem: the observed difference in mean outcomes does not identify the causal effect:

$$\underbrace{E(Y_i \mid D_i = 1)}_{=12} - \underbrace{E(Y_i \mid D_i = 0)}_{=6} = \underbrace{E(Y_i(1) - Y_i(0) \mid D_i = 1)}_{\text{ATT}} + \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{Selection bias} > 0 \text{ if risky roads get camera}}$$

- Sol. 1 (RCT): Randomize where cameras go among eligible locations.
  - » Identify 100 similar high-risk roads. Randomly assign: 50 roads get cameras, 50 don't.
  - » Compare crash rates after installation. Suppose we find that roads with cameras have on average 8 crashes/year, and roads without have 10. Then the estimated causal effect of the cameras is –2 crashes/year.

- Sol. 2 (selection on observables): Control for factors driving camera placement and crash risk.
  - » Compare camera roads to non-camera roads with similar risk factors (e.g. same traffic volume, speed limit and measured speeds, number of lanes, lighting, curvature...).
  - » Conditional on all the risk factors, observed mean difference identifies the causal effect.

Selection on observables

- Assume that, after controlling for observed covariates X_i, treated and untreated units have the same average potential outcomes. We have **conditional mean independence**:

$$E[Y_i(1) \mid D_i, X_i] = E[Y_i(1) \mid X_i], \qquad E[Y_i(0) \mid D_i, X_i] = E[Y_i(0) \mid X_i].$$

- Under this assumption, **the conditional selection bias becomes zero**:

$$E(Y_i(0) \mid D_i = 1, X_i) - E(Y_i(0) \mid D_i = 0, X_i) = E(Y_i(0) \mid X_i) - E(Y_i(0) \mid X_i) = 0$$

- So the conditional difference in mean outcomes identifies the conditional treatment effect:

$$E(Y_i \mid D_i = 1, X_i) - E(Y_i \mid D_i = 0, X_i) = E(Y_i(1) - Y_i(0) \mid X_i) = CATE(X_i)$$

and, taking the unconditional expectation:

$$E\big(E(Y_i(1) - Y_i(0) \mid X_i)\big) = E(Y_i(1) - Y_i(0)) = ATE$$

Example (cont'd)

- Let $X_i \in \{H, L\}$ denote road risk. Within each risk group, cameras reduce crashes:

	$D = 1$ (camera)	$D = 0$ (no camera)
$X = H$ (high risk)	$E(Y \mid D = 1, H) = 16$	$E(Y \mid D = 0, H) = 18$
$X = L$ (low risk)	$E(Y \mid D = 1, L) = 0$	$E(Y \mid D = 0, L) = 2$

  So $E(Y \mid D = 1, X) - E(Y \mid D = 0, X) = -2 < 0$ for both $X = H$ and $X = L$.

- **Selection into treatment**: cameras are more likely on high-risk roads:

$$P(H \mid D = 1) = 0.75, \qquad P(H \mid D = 0) = 0.25.$$

$$\Longrightarrow \qquad E(Y \mid D = 1) = 0.75 \cdot 16 + 0.25 \cdot 0 = 12, \qquad E(Y \mid D = 0) = 0.25 \cdot 18 + 0.75 \cdot 2 = 6.$$

- **Simpson's paradox**: cameras reduce crashes within each risk group, but the overall difference is positive:

$$E(Y \mid D = 1) - E(Y \mid D = 0) = 12 - 6 > 0,$$

  because treated roads contain a much larger share of high-risk locations (selection bias).

Recap: independence assumptions

	Unconditional	**Conditional on X**
Full independence	$(Y(1), Y(0)) \perp\!\!\!\perp D$ same distribution across D	$(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$ same distribution across D within each X
Mean independence	$E[Y(0) \mid D] = E[Y(0)]$ $E[Y(1) \mid D] = E[Y(1)]$ same means across D	$E[Y(0) \mid D, X] = E[Y(0) \mid X]$ $E[Y(1) \mid D, X] = E[Y(1) \mid X]$ same means across D within each X

- To remove selection bias in the mean, it is enough to assume mean independence.
- In an RCT, randomization gives full independence. This supports comparisons beyond averages (e.g. quantiles or tail probabilities), with minimal modeling.
- With non-experimental data we assume conditional mean independence given X; the main risk is *unobserved confounding* (omitted variables correlated with both D and Y).

Common support

- We have seen that:

$$CATE(X_i) = E(Y_i \mid D_i = 1, X_i) - E(Y_i \mid D_i = 0, X_i).$$

  So, to identify the conditional treatment effect, we need to observe both $E(Y_i \mid D_i = 1, X_i)$ and $E(Y_i \mid D_i = 0, X_i)$. This requires an additional key assumption.

- **Common support**: for all x in the support of X,

$$0 < P(D_i = 1 \mid X_i = x) < 1.$$

- Intuition: within each covariate group $X = x$, we need *both* treated and untreated units. Otherwise, we extrapolate counterfactual outcomes outside the data and the estimated effect is driven by modeling assumptions.

- Example: if all high-risk roads get cameras and no low-risk roads do, we cannot estimate the within-risk effect.

Propensity score

- If the dimension of X is large, estimating $E(Y_i(d) \mid D_i = d, X_i = x)$ is imprecise (few treated/controls within each X-cell).
- Solution: summarize all information in X into a one-dimensional variable, the **propensity score**:

$$\pi(x) = P(D_i = 1 \mid X_i = x).$$

- **Balancing property:** conditional on $\pi(X_i)$, covariates are balanced across treatment:

$$D_i \perp\!\!\!\perp X_i \mid \pi(X_i).$$

  Intuition: once you condition on the probability of treatment implied by X, knowing X gives no additional information about whether $D = 1$.

- If selection on observables holds given X_i, then it also holds given the propensity score. Hence, controlling for $\pi(X_i)$ is enough to remove selection bias.
- For identification we still need common support: $0 < \pi(X_i) < 1$.

Application

- There is a (fictional) government program aiming at reducing the use of pollutant fertilizer in carrot production by subsidizing the supply of organic fertilizers (minerals).
- To investigate whether the program is effective, a pilot is run in 5 French regions: a fixed quantity of minerals per hectare is given to the first 200 farms that apply.
- In the dataset: farm's identifier (`id`), characteristics (`size`, `protected`, ...), and use of pollutant fertilizer (`fertilizer_2018`, and `fertilizer_2020`);
  the dummy variable `eligible` equals 1 if the farm is located in one of the eligible regions;
  the dummy variable `d` equals 1 if the farmer receives the minerals.
- Questions:
  - » *What is the unit of observation?*
  - » *What are the treatment and outcome variables?*
  - » *Is there randomization?*
  - » *Is there selection on observables?*
  - » *Could we have selection bias?*

Application (cont'd)

- We work with **simulated data**: we know the data generating process, including the effect of the covariates on the outcome, and we measure how far our estimates are from the true coefficients.
- In particular, we know the ATE: $E(Y_i(1) - Y_i(0)) = -10$ kg/Ha.
- We can try regressing:

$$Y_i = \alpha + \beta D_i + \epsilon_i.$$

The coefficient $\hat{\beta}$ is an estimate of

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = ATT + \text{ selection bias.}$$

Remember: selection bias $= E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0)$.

Application (cont'd)

- Alternatively, try controlling for X_i:

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \epsilon_i.$$

Assume:

  - » **selection on observables**: conditional selection bias is zero and conditional difference in mean outcomes identifies the conditional treatment effect;
  - » constant treatment effect: $CATE(X_i) = \tau, \ \forall X_i$;
  - » correct specification for the X part.

Then the OLS coefficient $\hat{\beta}$ **recovers the ATE**.

- You can also use the **propensity score**:

$$Y_i = \alpha + \beta D_i + f\left(\hat{\pi}(X_i)\right) + \epsilon_i.$$

  - » Useful when X is high-dimensional,
  - » but can reduce precision and is sensitive to how we model $f(\cdot)$, $\hat{\pi}(X_i)$ and to common support.