Notes on Lab Session 2

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

TSE - MSc course in Program Evaluation
January 2026

Definitions

- **Treatment** $D \in \{0, 1\}$: observed variable whose impact is to be measured
  Outcome Y: observed variable on which the impact is measured
- The impact of the treatment on observation i is:

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

  Key issue: it is impossible to observe *at the same time* the outcome if i takes the treatment ($Y_i(1)$) *and* the outcome if i does not take the treatment ($Y_i(0)$)
- Idea: use multiple or repeated observations in which some are treated and some are untreated and compare the treated group to the control group
- Average Treatment Effect (**ATE**):

$$ATE = E(Y_i(1) - Y_i(0))$$

- Average Treatment Effect on the Treated (**ATT**):

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

The key issue of observability

- Assuming SUTVA, we can rewrite observed response Y_i as a function of potential outcomes:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

- The difference that we observe is:

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

$$= \underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1)}_{\text{Average Treatment Effect on the Treated}} + \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{Selection bias}}$$

- The **selection bias** captures baseline differences between treated and controls: treated and untreated units would have had different outcomes *even without treatment*

  $\longrightarrow$ Ex: those selected into job training programs might earn lower incomes because of lower skills

- A randomized control trial ensures that there is no selection bias

Randomized control trial

- In a randomized experiment, the treatment variable D_i is chosen at random:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i$$

- This implies mean independence:

$$E(Y_i(1) \mid D_i = 1) = E(Y_i(1) \mid D_i = 0) = E(Y_i(1))$$

$$E(Y_i(0) \mid D_i = 1) = E(Y_i(0) \mid D_i = 0) = E(Y_i(0))$$

  Expected potential outcomes are the same for treatment and control groups

- Hence **ATT is equal to ATE**:

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

  and there is **no selection bias**:

$$E(Y_i(0) \mid D_i = 1) = E(Y_i(0) \mid D_i = 0)$$

Estimation

- **Group mean differences**

  In an RCT $ATT = ATE = E(Y_i(1)) - E(Y_i(0))$ can be estimated by its sample analogue:

$$\widehat{ATE} = \frac{1}{n_T}\sum_{i \in T} Y_i - \frac{1}{n_C}\sum_{i \in C} Y_i$$

  where T is treatment group and C is control group

- **Linear regression**

$$Y_i = \alpha_0 + \alpha_1 D_i + \varepsilon_i \quad \Longleftrightarrow \quad \begin{cases} Y_i = \alpha_0 + \varepsilon_i & \text{if } D_i = 0 \\ Y_i = \alpha_0 + \alpha_1 + \varepsilon_i & \text{if } D_i = 1 \end{cases}$$

$$\min_{\alpha_0, \alpha_1} \sum_i (Y_i - \alpha_0 - \alpha_1 D_i)^2 = \min_{\alpha_0} \sum_{i \in C} (Y_i - \alpha_0)^2 + \min_{\alpha_0 + \alpha_1} \sum_{i \in T} (Y_i - (\alpha_0 + \alpha_1))^2$$

$$\hat{\alpha}_0 = \frac{1}{n_C}\sum_{i \in C} Y_i, \qquad \widehat{(\alpha_0 + \alpha_1)} = \frac{1}{n_T}\sum_{i \in T} Y_i \;\Rightarrow\; \hat{\alpha}_1 = \frac{1}{n_T}\sum_{i \in T} Y_i - \frac{1}{n_C}\sum_{i \in C} Y_i = \widehat{ATE}$$

Randomization

- Randomization can occur at different levels:
  - » individual level: each person is randomly assigned to treatment/control
  - » coarser levels: groups of individuals (e.g. a village in Progresa) are randomized together, so everyone in the group either receives the treatment or not
- Individual-level randomization is ideal but may be infeasible because of:
  - » inability to control individual access to treatment
  - » risk of contagion/spillovers from treated to control units
- **In finite samples**, randomization can still create chance **imbalance** in key covariates: if they strongly affect outcomes, treated and control groups may differ in baseline $Y(0)$ on average
  $\longrightarrow$ differences in outcome may reflect differences in group composition, not treatment
- **Stratified randomization** balances treatment and control groups on these key covariates:
  1) identify important covariates (e.g. gender, education)
  2) create blocks for each covariate combination (e.g. men without a high-school degree)
  3) randomly assign individuals *within each block* to treatment/control

Role of controls in RCTs

Adding controls in an RCT plays no role in terms of identification of the ATE, but is useful to:

- Increase precision in the estimation
- Test for validity of randomization
  - » For any baseline covariates X_i, we have independence by construction:

$$X_i \perp\!\!\!\perp D_i \implies E(X_i \mid D_i = 1) = E(X_i \mid D_i = 0)$$

    So we can check the equality of means in the treatment and control samples

  - » Also, covariates should not predict treatment D_i (no joint significant effect)

$$D_i = \delta_0 + X_i'\delta + u_i, \qquad H_0 : \ \delta = 0$$

Application: the NSW program

- The NSW program (U.S., 1970s) provided job training and subsidized employment to disadvantaged individuals.
- Eligible individuals were recruited, and then randomly assigned to a treatment group offered the program or to a control group not offered the program.
- The goal was to estimate whether the program improved subsequent labor-market outcomes.

- Question: how much more/less did a person earn after being assigned to the NSW program relative to not being assigned?
- Unit of observation: individual (person belonging to eligible population)
- Randomization level: individual
- Treatment variable D_i: whether the person was assigned or not to the program (`treat`)
- Outcome variable Y_i: post-program earnings of the person (`re78`)
- Covariates X_i: baseline characteristics (`age`, `education`, ...)