Chapter 6: Policy Evaluation Methods: Treatment Effects

Joan Llull

Microeconometrics
IDEA Phd Program

Potential Outcomes and Causality

Potential Outcomes

Consider the **population** of individuals susceptible of a treatment:

- Y_{1i}: outcome for individual i if exposed to the treatment ($D_i = 1$)
- Y_{0i} be the outcome for the same individual if not exposed ($D_i = 0$)
- Treatment indicator: D_i

Note that Y_{1i} and Y_{0i} are **potential outcomes** in the sense that we only observe:

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i).$$

Main challenge of this approach: the treatment effect can not be computed for a given individual.

Our interest is not in treatment effects for **specific individuals** *per se*, but, instead, in some characteristics of their distribution.

Treatment Effects

Most of the time focus on two main parameters of interest:

The first one is the **average treatment effect** (ATE):

$$\alpha_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}],$$

The second is **average treatment effect on the treated** (TT):

$$\alpha_{TT} \equiv \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1].$$

As noted, the main challenge is that we **only observe** Y_i. The standard measure of association between Y_i and D_i is:

$$\begin{aligned}\beta &\equiv \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]}_{\alpha_{TT}} + \underbrace{(\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0])}_{\text{selection bias}}.\end{aligned}$$

which differs from α_{TT} unless the second term is equal to zero.

The second term (selection bias) indicates the difference in potential outcomes when **untreated for individuals** that are actually treated and individuals that are not.

A nonzero difference may result from a situation in which treatment status is the result of individual decisions where those with low Y_0 choose treatment more frequently than those with high Y_0 (**difference in composition**).

An important assumption of the potential outcome representation is that the effect of the treatment on one individual is **independent of the treatment received by other** individuals. This excludes equilibrium or feedback effects, as well as strategic interactions among agents.

Structural vs Reduced/Form Effects

From a **structural model** of D_i and Y_i, one could obtain the implied average treatment effects.

Instead, here, they are defined with respect to the distribution of potential outcomes, so that, relative to the structure, they are **reduced-form causal effects**.

Econometrics has conventionally distinguished between **reduced form** effects, uninterpretable but useful for prediction, and **structural** effects, associated with rules of behavior.

The treatment effects provide this **intermediate category** between predictive and structural effects, in the sense that recovered parameters are causal effects, but they are uninterpretable in the same sense as reduced form effects.

Sample Counterparts

The sample average version of β is given by:

$$\begin{aligned}\beta^S &\equiv \bar{Y}_T - \bar{Y}_C \\ &\equiv \frac{1}{N_1}\sum_{\mathrm{i}=1}^{\mathrm{N}} Y_i D_i - \frac{1}{N_0}\sum_{\mathrm{i}=1}^{\mathrm{N}}(1-D_i)Y_i,\end{aligned}$$

where $N_0 \equiv N - N_1$ is the number of untreated individuals.

Identification: Independence

Identification of the treatment effects depends on the **assumptions** we make on the relation between potential outcomes and the treatment.

Simplest case is when the distribution of the potential outcomes is **independent** of the treatment (e.g. randomized experiments):

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i.$$

When this happens:

$$F(Y_{1i}|D_i = 1) = F(Y_{1i})$$
$$F(Y_{0i}|D_i = 0) = F(Y_{0i})$$

which implies that:

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

and, as a result, $\alpha_{ATE} = \alpha_{TT} = \beta \Rightarrow$ **unbiased estimate** of α_{ATE}:

$$\widehat{\alpha}_{ATE} = \bar{Y}_T - \bar{Y}_C = \beta^S.$$

No need to "control" for other **covariates**.

Identification: Conditional Independence

A less restrictive assumption is **conditional independence**:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i,$$

where X is a vector of covariates.

This situation is known as **matching**: for each "type" of individual (i.e. each value of covariates) we match treated and controls.

Conditional independence implies:

$$\mathbb{E}[Y_{ji}|X] = \mathbb{E}[Y_{ji}|D_i = j, X_i] = \mathbb{E}[Y_i|D_i = j, X_i] \text{ for } j = 0, 1$$

and, as a result:

$$\alpha_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}] = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i),$$

For the treatment effect on the treated:

$$\begin{aligned} \alpha_{TT} &= \int \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1, X_i] dF(X_i|D_i = 1) \\ &= \int \mathbb{E}[Y_i - \mu_0(X_i)|D_i = 1, X_i] dF(X_i|D_i = 1), \end{aligned}$$

where $\mu_0(X_i) \equiv \mathbb{E}[Y_i|D_i = 0, X_i]$. The function $\mu_0(X_i)$ is used as an imputation for Y_{0i}.

Identification: Absence of Independence

Finally, sometimes we cannot assume conditional independence:

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i | X_i.$$

In this case, we will need some variable Z_i that constitutes an **exogenous** source of variation in D_i, in the sense that it satisfies the **independence assumption**:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i | X_i,$$

and the **relevance condition**:

$$Z_i \not\perp\!\!\!\perp D_i | X_i.$$

As we discuss later in the course, in this context we are only able to identify an average treatment effect for a subgroup of individuals, known as **local average treatment effect**.

Randomized Control Trials and Natural Experiments

Randomized Experiments

In the treatment effect approach, a **randomized field trial** is regarded as the ideal research design.

Long **history** of randomized field trials in social welfare in the U.S., beginning in the 1960s (see Moffitt (2003) for a review).

Encouraged by U.S. Federal Government, eventually almost mandatory. Legislation introduced in 1988.

Resistance from many states on ethical grounds (more so in other countries, where treatment groups are often areas for treatment instead of individuals).

Sometimes experiments are provided by nature: **natural experiments** (e.g. John Snow and the cholera case in SoHo).

Random Assignment and Treatment Effects

In a controlled experiment, treatment status is **randomly assigned** by the researcher, which by construction, ensures **independence**:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i.$$

As noted, this **eliminates the selection bias** as:

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

Also $\alpha_{ATE} = \alpha_{TT} = \beta$, as $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i}]$.

Thus, the average treatment effect can be estimated by a simple **linear regression** of the observed outcome Y_i on the treatment dummy D_i and a constant.

Introduction of Additional Regressors

Additional regressors W_i are not needed for **consistency** as:

$$\gamma \frac{\text{Cov}(W_i, D_i)}{\text{Var}(D_i)} = 0.$$

Yet, it can be interesting to include them for several reasons:

- If they are relevant, they can increase **precision** (Frisch-Waugh Theorem).
- **Checking randomization**: are there statistical difference in these regressors between treated and controls?
- **Used in the randomization** (e.g. village-level randomization).

The last two lead to the context of **conditional independence**.

Warning: Partial Compliance

So far we have assumed **perfect complience**: everyone elected takes the treatment and no control takes it.

Now: $D_i = \mathbb{1}\{\text{treatment taken}\}$ and $Z_i = \mathbb{1}\{\text{assigned to treatment}\}$.

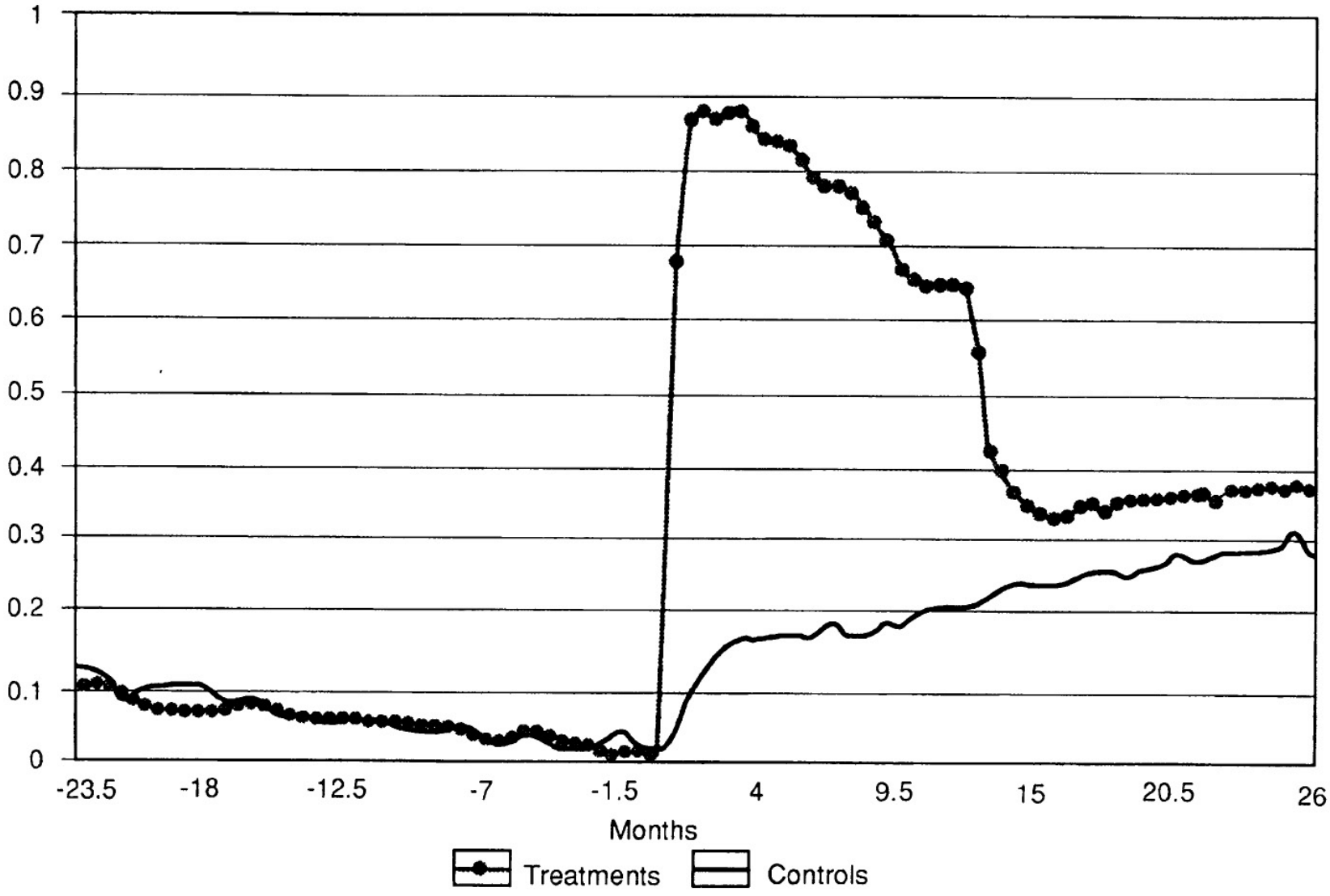
We may have $D_i = 0$ and $Z_i = 1$ (**no-shows**), and $D_i = 1$ and $Z_i = 0$ (**cross-overs**).

Now $Y_{1i}, Y_{0i} \not\perp\!\!\!\perp D_i$ but $Y_{1i}, Y_{0i} \perp\!\!\!\perp Z_i$. The latter can be used in an **IV fashion** to obtain α_{TT} (see IV section below), or compute an **intention-to-treat** effect.

Warning: Longer Run Outcomes

National Supported Work program (NSW):

- designed in the U.S. in the mid 1970s
- training and job opportunities to disadvantaged workers
- NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards.
- experimental design on women who volunteered for training
- Requirements: unemployed, a long-term AFDC recipient, and have no preschool children
- Participants were randomly assigned to treatment (275) and control groups (266) in 1976-1977
- Training in 1976, and then followed.
- Ham and LaLonde (1996) analyze the effects of the program.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-23.5
-18
-12.5
-7
-1.5
4
9.5
15
20.5
26
Months
Treatments
Controls

Effects on Unemployment Rates

Thanks to randomization, comparison between employment rates of treatments and controls gives an **unbiased estimate** of the effect of the program on employment at different horizons.

Initially, by construction there is a mechanical effect from the fact that treated women are offered a **subsidized job**.

Compliance with the treatment is decreasing over time, as women can decide to **drop from the subsidized job**.

The **employment growth for controls** is just a reflection of the program's eligibility criteria.

Importantly, after the program ends, a **9 percentage points difference** in employment rates is sustained.

Ham and LaLonde's Additional Point

But Ham and LaLonde (1996) make an important additional point: randomization **does not guarantee independence** for any possible outcomes.

Two examples: wages and unemployment durations (hazards).

Effect of training program on employment rates of the treated $\Rightarrow$ those who are working are a **selected sample**.

Notation: W_i wages; $Y_i = 1$ if employed; $\eta_i = 1$ skilled type.

Suppose:

$$P(Y_i = 1|D_i = 1, \eta_i = j) > P(Y_i = 1|D_i = 0, \eta_i = j), \quad j = 0, 1$$

and:

$$\frac{P(Y_i = 1|D_i = 1, \eta_i = 0)}{P(Y_i = 1|D_i = 0, \eta_i = 0)} > \frac{P(Y_i = 1|D_i = 1, \eta_i = 1)}{P(Y_i = 1|D_i = 0, \eta_i = 1)}.$$

This implies that the **frequency of low skill** will be greater in the group of employed treatments than in the employed controls:

$$P(\eta_i = 0|Y_i = 1, D_i = 1) > P(\eta_i = 0|Y_i = 1, D_i = 0),$$

which is a way to say that η_i, which is unobserved, is **not independent** of D_i given $Y_i = 1$, although, unconditionally, $\eta_i \perp\!\!\!\perp D_i$.

Consider the **conditional effects**:

$$\Delta_j \equiv \mathbb{E}[W|Y_i = 1, D_i = 1, \eta_i = j] - \mathbb{E}[W_i|Y_i = 1, D_i = 0, \eta_i = j], \quad j = 0, 1$$

Our effect of interest is:

$$\Delta_{ATE} = \Delta_0 P(\eta_i = 0) + \Delta_1 P(\eta_i = 1),$$

and comparison of average wages between treatments and controls is:

$$\Delta_W = \mathbb{E}[W_i|Y_i = 1, D_i = 1] - \mathbb{E}[W_i|Y_i = 1, D_i = 0] < \Delta_{ATE}.$$

$\Rightarrow$ may not be possible to correctly measure the effect on wages.

Matching

Selection Based on Observables and Matching

Experiments are often **too expensive**, **unfeasible**, or **unethical** (e.g. smoking on mortality), and sometimes randomization on observables $\Rightarrow$ observational data (unlikely to satisfy independence).

In many situations, we can defend **conditional independence**:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | X_i.$$

As we saw before:

$$\alpha_{ATE} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i),$$

$$\alpha_{TT} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i|D_i = 1).$$

Matching: compares individuals with the same characteristics and then integrates over the distribution of characteristics.

The common support condition

Essential condition for matching: for each possible value of X, there are individuals in the treatment and control group for which we can average outcomes $\Rightarrow$ **common support condition**:

$$0 < P(D_i = 1|X_i) < 1 \quad \text{for all } X_i \text{ in its support.}$$

Counterexample (with a single covariate):

$$P(D_i = 1|X_i) = \begin{cases} 1 & \text{if } X_{min} \leq X < \underline{X} \\ p \in (0,1) & \text{if } \underline{X} \leq X \leq \overline{X} \\ 0 & \text{if } \overline{X} < X \leq X_{max} \end{cases}.$$

Implication:

- $\mathbb{E}[Y_i|D_i = 1, X_i]$ only identified for values of X_i in $(X_{min}, \overline{X})$
- $\mathbb{E}[Y_i|D_i = 0, X_i]$ only identified for values of X_i in $(\underline{X}, X_{max})$
- $\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$ only for values of X_i in the intersection range $(\underline{X}, \overline{X}) \Rightarrow \alpha_{ATE}$ and α_{TT} not identified

Propensity Score Matching

Sometimes, set of variables X_i is too large or multivariate.

Not all info in X_i is relevant $\Rightarrow$ **propensity score matching**.

Rosenbaum and Rubin (1983) defined the **propensity score** as:

$$\pi(X_i) \equiv P(D_i = 1|X_i).$$

and note it is a **sufficient statistic** for the distribution of D_i.

Thus:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|X_i \quad \Leftrightarrow \quad Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|\pi(X_i).$$

$\Rightarrow$ match on the propensity score instead of the covariates.

Two-step methods: estimate the propensity score, and then create the appropriate weighting.

Under (unconditional) **independence** we established that:

$$\alpha_{ATE} = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \frac{\mathbb{E}[D_i Y_i]}{P(D_i = 1)} - \frac{\mathbb{E}[(1 - D_i)Y_i]}{P(D_i = 0)}.$$

Thus, under **conditional independence** we can write:

$$\begin{aligned}\mathbb{E}[Y_{1i} - Y_{0i}|X_i] &= \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i] \\ &= \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)(1 - \pi(X_i))}\middle| X_i\right],\end{aligned}$$

and:

$$\alpha_{ATE} = \mathbb{E}\left[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]\right] = \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)[1 - \pi(X_i)]}\right].$$

Estimation: Discrete Low-Dimensional Case

Notation:

- X_i is discrete and takes on J possible values $\{x_j\}_{j=1}^J$
- N observations $\{X_i\}_{i=1}^N$
- N^j is the number of observations in cell j
- N_ℓ^j be the number of observations in cell j with $D_i = \ell$
- $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D_i = \ell$

Note $\bar{Y}_1^j - \bar{Y}_0^j$ is the sample counterpart of $\mathbb{E}[Y_i|D_i = 1, X_i = x_j] - \mathbb{E}[Y_i|D_i = 0, X_i = x_j]$, which can be used to get the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N^j}{N}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N_1^j}{N_1} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i - \bar{Y}_0^{j(i)}\right).$$

where $j(i)$ indicates the cell of X_i (note matching interpretation of the second expression for $\widehat{\alpha}_{TT}$).

Estimation: Propensity Score Weighting

Using the sample analog of α_{ATE} in terms of the propensity score (**Hirano, Imbens, and Ridder, 2003**):

$$\widehat{\alpha}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \left(\frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1 - \widehat{\pi}(X_i)]} \right),$$

where $\widehat{\pi}(X_i)$ is obtained in a **first stage** either nonparametrically, or by means of a flexibly specified Logit or Probit.

Estimation Methods: Weigthing

A **matching estimator** can be regarded as a way of constructing **imputations** for missing potential outcomes in a similar way, so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit.

In the **exact matching** we were doing:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k:D_k=0} Y_k \frac{\mathbb{1}\{X_k = X_i\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{X_\ell = X_i\}}.$$

More generally we can compute:

$$\widehat{Y}_{0i} = \sum_{k:D_k=0} w(i,k) Y_k,$$

where **different weighting** $w(i,k)$ determine different estimators.

- **Nearest neighbor matching** (with replacement):

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\},$$

(picking the individual k in the control group with the closest observables to the individual i in the treated group).

- **Radius matching** (with replacement):

$$w(i,k) = \frac{\mathbb{1}\{||X_k - X_i|| < \varepsilon\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{||X_\ell - X_i|| < \varepsilon\}},$$

for some threshold ε (averages the observations from the control group with covariates within a window centered at X_i).

- **Kernel matching** (with replacement):

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell:D_\ell=0} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)},$$

where $\kappa(\cdot)$ is a kernel function that downweights distant observations, and γ_{N_0} is a bandwidth parameter.

They can also be used for the **propensity score** $\pi(X_i)$ rather than X_i.

Instrumental Variables

Identification in IV Settings

Suppose:

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i | X_i,$$

but there is some variable Z_i that satisfies **independence condition**:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i | X_i,$$

and the **relevance condition**:

$$Z_i \not\perp\!\!\!\perp D_i | X_i.$$

Matching can be regarded as a special case in which $Z_i = D_i$, i.e. all the variation in D_i is exogenous given X_i.

For simplicity, we do most of the analysis below considering a **single binary instrument** Z_i, and we abstract from including **covariates**.

Two cases: homogeneous and heterogeneous treatment effects.

Homogeneous Treatment Effects

In this case, the causal effect is the **same for every individual**:

$$Y_{1i} - Y_{0i} = \alpha \quad \forall i.$$

Availability of an instrumental variable allows us to **identify** α (traditional situation in econometric models — IV regression):

$$\alpha = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, D_i)}.$$

Alternatively:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = Y_{0i} + \alpha D_i.$$

Taking into account that $Y_{0i} \perp\!\!\!\perp Z_i$ (conditional independence and subtract):

$$\alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]},$$

Identification **requires** $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] \neq 0$ (relevance).

We get the effect of D_i on Y_i through the effect of Z_i because Z_i only affects Y_i through D_i (**exclusion restriction**).

Heterogeneous Treatment Effects

In the heterogeneous case, the availability of instrumental variables is **not sufficient** to identify a causal effect (e.g. α_{ATE}).

Monotonicity condition: any person that was willing to treat if assigned to the control group would also be prepared to treat if assigned to the treatment group.

The **plausibility** of this assumption depends on the context of the application.

Under monotonicity, the IV coefficient coincides with the average treatment effect for those whose value of D_i would change when changing the value of Z_i, which is known as the **local average treatment effect** (LATE).

Potential Treatment Representation:

Let:

- D_{0i}: D_i when $Z_i = 0$
- D_{1i}: D_i when $Z_i = 1$

Only observe $D_{\ell i}$, for ℓ either equal to one or to zero $\Rightarrow$ **four observable** groups, eight **potential groups**:

Obs. type	Z_i	D_i	D_{0i}	D_{1i}	**Latent type**
Type 1	0	0	0	0	Never-taker
				1	Complier
Type 2	0	1	1	0	Defier
				1	Always-taker
Type 3	1	0	0	0	Never-taker
			1		Defier
Type 4	1	1	0	1	Complier
			1		Always-taker

Role of monotonicity:

Now we have:

$$\mathbb{E}[Y_i|Z_i = 1] = \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i} - Y_{0i})D_{1i}]$$
$$\mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i} - Y_{0i})D_{0i}],$$

which implies:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] &= \mathbb{E}[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} - D_{0i} = 1]P(D_{1i} - D_{0i} = 1) \\ &\quad - \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} - D_{0i} = -1]P(D_{1i} - D_{0i} = -1).\end{aligned}$$

Thus, $\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$ **could be negative** and yet the causal effect be **positive for everyone**, as long as the probability of *defiers* is sufficiently large.

Imperfect Compliance and IV

Stronger assumption than monotonicity:

$$P(D_i = 1|Z_i = 0) = 0,$$

(no treatment for individuals with $Z_i = 0$) $\Rightarrow$ **imperfect compliance**.

In this case:

$$\mathbb{E}[Y_i|Z_i = 1] = \mathbb{E}[Y_{0i}] + \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 1]P(D_i = 1|Z_i = 1),$$

and, since $P(D_i = 1|Z_i = 0) = 0$:

$$\mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_{0i}].$$

Therefore:

$$\alpha_{TT} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{P(D_i = 1|Z_i = 1)}$$

(where we use $P(Z_i = 1|D_i = 1) = 1$.

$\Rightarrow$ if the eligibility condition holds, the **IV coefficient** coincides with the **treatment effect on the treated**.

Local Average Treatment Effects (LATE)

Ruling out defiers (which implies monotonicity):

$$\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]=\mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1]P(D_{1i}-D_{0i}=1),$$

$$\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]=\mathbb{E}[D_{1i}-D_{0i}]=P(D_{1i}-D_{0i}=1).$$

Local average treatment effect (LATE):

$$\alpha_{LATE}\equiv\mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1]=\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}.$$

Imbens and Angrist (1994) called like this because it is the average treatment effects **on the subsample of compliers**.

$\Rightarrow$ different instrumental variables lead to **different parameters**, even under instrument validity, which is counter to standard GMM thinking.

$\Rightarrow$ need to think of the **group of compliers selected** by the instrument (policy relevant instruments).

This concept changed radically the **way we think** of and understand IV.

Relevance requires presence of compliers.

Conditional Estimation with IV

Assume independence and relevance only cold **conditionally**:

$$\begin{array}{ll} (Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i | X_i & \text{(conditional independence)} \\ Z_i \not\perp\!\!\!\perp D_i | X_i & \text{(conditional relevance)} \ . \end{array}$$

Example: distance to college, Z_i, is not randomly assigned but chosen by parents, and this choice may depend on family background, X_i.

In general, we now have a **conditional LATE** given X_i:

$$\gamma(X_i) \equiv \mathbb{E}[Y_{1i} - Y_{0i} | D_{1i} - D_{0i} = 1, X_i],$$

and a conditional IV estimator:

$$\beta(X_i) \equiv \frac{\mathbb{E}[Y_i | Z_i = 1, X_i] - \mathbb{E}[Y_i | Z_i = 0, X_i]}{\mathbb{E}[D_i | Z_i = 1, X_i] - \mathbb{E}[D_i | Z_i = 0, X_i]}.$$

Aggregate effect: we proceed differently depending on whether the effects are homogeneous or heterogeneous.

In the **homogeneous** case:

$$Y_{1i} - Y_{0i} = \beta(X_i) \quad \forall i.$$

In the **heterogeneous** case, it makes sense to consider an average treatment effect for the **overall subpopulation of compliers**:

$$\begin{aligned}\beta_C &\equiv \int \beta(X_i)\frac{P(compliers|X_i)}{P(compliers)}dF(X_i) \\ &= \int \{\mathbb{E}[Y_i|Z_i=1,X_i]-\mathbb{E}[Y_i|Z_i=0,X_i]\}\,\frac{1}{P(compliers)}dF(X_i),\end{aligned}$$

where:

$$P(compliers) = \int \{\mathbb{E}[D_i|Z_i=1,X_i]-\mathbb{E}[D_i|Z_i=0,X_i]\}\,dF(X_i).$$

Therefore:

$$\beta_C = \frac{\int \{\mathbb{E}[Y_i|Z_i=1,X_i]-\mathbb{E}[Y_i|Z_i=0,X_i]\}\,dF(X_i)}{\int \{\mathbb{E}[D_i|Z_i=1,X_i]-\mathbb{E}[D_i|Z_i=0,X_i]\}\,dF(X_i)},$$

which can be estimated as a **ratio of matching estimators** (Frölich, 2003).

Regression Discontinuity

The Fundamental RD Assumption

In regression discontinuity we consider a situation where there is a **continuous** variable Z that is not necessarily a valid instrument (it does not satisfy the exogeneity assumption), but such that **treatment** assignment is a **discontinuous function** of Z:

$$\lim_{z \to z_0^+} P(D_i = 1 | Z_i = z) \neq \lim_{z \to z_0^-} P(D_i = 1 | Z_i = z)$$

$$\lim_{z \to z_0^+} P(Y_{ji} \leq r | Z_i = z) = \lim_{z \to z_0^-} P(Y_{ji} \leq r | Z_i = z) \quad (j = 0, 1)$$

which are **relevance** and **orthogonality** conditions respectively.

Implicit regularity conditions are:

- existence of the limits,
- Z_i has positive density in a neighborhood of z_0.

For now we abstract from **conditioning covariates** for simplicity.

Sharp and Fuzzy Designs

Early RD literature in Psychology (Cook and Campbell, 1979) distinguishes between:

- **Sharp design**: $D_i = \mathbb{1}\{Z_i \geq z_0\}$, with:

$$\lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] = 1$$
$$\lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z] = 0.$$

- **Fuzzy design**: $0 < P(D_i = 1|Z_i \geq z_0) < 1$, with:

$$P(D_i = 1|Z_i = z_0 - \varepsilon) \neq P(D_i = 1|Z_i = z_0 + \varepsilon)$$

Homogeneous Treatment Effects

Suppose that $\alpha_i = Y_{1i} - Y_{0i}$ is **constant**, so that $Y_i = \alpha D_i + Y_{0i}$.

Conditional expectations given $Z_i = z$ and left- and right-side limits:

$$\lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] = \alpha \lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_{0i}|Z_i = z]$$
$$\lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z] = \alpha \lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_{0i}|Z_i = z],$$

which leads to the consideration of the following **RD parameter**:

$$\alpha = \frac{\lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z]}{\lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z]}.$$

determined by **relevance** and **orthogonality** conditions above.

In the case of a sharp design, the denominator is unity so that:

$$\alpha = \lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z],$$

Sharp corresponds to **matching** and fuzzy corresponds to **IV**.

Intuitively, considering units within a small interval around the cutoff point is similar to a **randomized experiment** at the cutoff point.

Heterogeneous Treatment Effects: Sharp

Now suppose that: $Y_i = \alpha_i D_i + Y_{0i}$.

In the **sharp** design since $D_i = \mathbb{1}\{Z_i \geq z_0\}$ we have:

$$\mathbb{E}[Y_i|Z_i = z] = \mathbb{E}[\alpha_i|Z_i = z]\,\mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_{0i}|Z = z].$$

Average treatment effect for individuals **at the threshold** value z_0:

$$\alpha_{RD} \equiv \mathbb{E}[\alpha_i|Z_i = z_0].$$

Thus, we can rewrite the above expression as:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i = z] &= \alpha_{RD}\,\mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_{0i}|Z_i = z] \\ &\quad + (\mathbb{E}[\alpha_i|Z_i = z] - \mathbb{E}[\alpha_i|Z_i = z_0])\,\mathbb{1}\{z \geq z_0\} \\ &\equiv \alpha_{RD} D_i + k_{z_0}(z).\end{aligned}$$

$\Rightarrow$ the situation is one of **selection on observables**.

Control function approach: the OLS population coefficient on D_i in the equation:

$$Y = \alpha_{RD} D + k(z) + w$$

equals $\mathbb{E}[\alpha_i|Z_i = z_0]$.

Heterogeneous Treatment Effects: Fuzzy

In the **fuzzy design**, D_i not only depends on $\mathbb{1}\{Z_i \geq z_0\}$, but also on other unobserved variables. Thus, D_i is an endogenous variable in the above regression.

We can use $\mathbb{1}\{Z_i \geq z_0\}$ as an **instrument** for D_i in such equation to identify α_{RD}, at least in the homogeneous case (connection with IV was first made explicit by van der Klaaw (2002)).

Below we discuss two alternative assumptions we can make for identification fuzzy designs: **conditional independence** near z_0, and **monotonicity**.

Conditional independence near z_0:

Weak conditional independence: $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | Z_i = z$ for z near z_0, i.e. for $z = z_0 \pm e$, where e is arbitrarily small positive number, or:

$$F(Y_{ji}|D_i = 1, Z_i = z_0 \pm e) = F(Y_{ji}|Z_i = z_0 \pm e) \quad (j = 0, 1).$$

An **implication** is:

$$\mathbb{E}[\alpha_i D_i | Z_i = z_0 \pm e] = \mathbb{E}[\alpha_i | Z_i = z_0 \pm e]\, \mathbb{E}[D_i | Z_i = z_0 \pm e].$$

Proceeding as before, we have:

$$\lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] = \lim_{z \to z_0^+} \mathbb{E}[\alpha_i|Z_i = z]\, \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_{0i}|Z_i = z]$$
$$\lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha_i|Z_i = z]\, \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_{0i}|Z_i = z].$$

Noting that $\lim_{z \to z_0^+} \mathbb{E}[\alpha_i|Z_i = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha_i|Z_i = z] = \alpha_{RD}$:

$$\alpha_{RD} \equiv \mathbb{E}[Y_{1i} - Y_{0i}|Z_i = z_0] = \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z]}.$$

That is, the RD parameter can be interpreted as the **average TE** at z_0.

Monotonicity near z_0:

Alternative assumption: **local monotonicity** (Hahn et al., 2001):

$$D_{z_0+\varepsilon,i} \geq D_{z_0-\varepsilon,i} \text{ for all units } i \text{ in the population,}$$

for some $\bar{\varepsilon} > 0$ and any pair $(z_0 - \varepsilon, z_0 + \varepsilon)$ with $0 < \varepsilon < \bar{\varepsilon}$, where D_{zi} is the potential assignment indicator associated with $Z_i = z$.

In some situations, **conditional independence** can be problematic and **local monotonicity** not.

In such cases, it can be shown that α_{RD} identifies the **local average treatment effect** at $z = z_0$:

$$\alpha_{RD} = \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{1i} - Y_{0i} | D_{z_0+\varepsilon,i} - D_{z_0-\varepsilon,i} = 1]$$

that is, the ATE for the units for whom treatment changes discontinuously at z_0.

Estimation Strategies

Hahn et al. (2001): Let $S_i \equiv \mathbb{1}\{z_0 - h < Z_i < z_0 + h\}$ where $h > 0$ denotes the bandwidth, and consider the subsample such that $S_i = 1$, and define $W_i \equiv \mathbb{1}\{z_0 < Z_i < z_0 + h\}$ as an instrument, applied to the subsample with $S_i = 1$:

$$\widehat{\alpha}_{RD} = \frac{\widehat{\mathbb{E}}[Y_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[Y_i|W_i = 0, S_i = 1]}{\widehat{\mathbb{E}}[D_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[D_i|W_i = 0, S_i = 1]}.$$

Alternative by the same authors, **control function**:

- **Sharp design**: OLS on $Y_i = \alpha_{RD} D_i + k(Z_i) + w_i$
- **Fuzzy design**: IV on $Y_i = \alpha_{RD} D_i + k(Z_i) + w_i$ using $\mathbb{1}\{Z_i \geq z_0\}$ as the excluded instrument.

Semiparametric approach (van der Klaaw, 2002): power series approximation for $k(Z)$.

The latter methods of estimation, not local to data points near the threshold, are implicitly predicated on the assumption of **homogeneous TE**.

Conditioning on Covariates

Even if the RD assumption is satisfied unconditionally, conditioning on covariates may **mitigate the heterogeneity in treatment effects**, hence contributing to the relevance of RD estimated parameters, which otherwise are "very local".

Covariates may also make the **local conditional exogeneity** assumption more credible.

Difference in Differences

Difference in differences setup

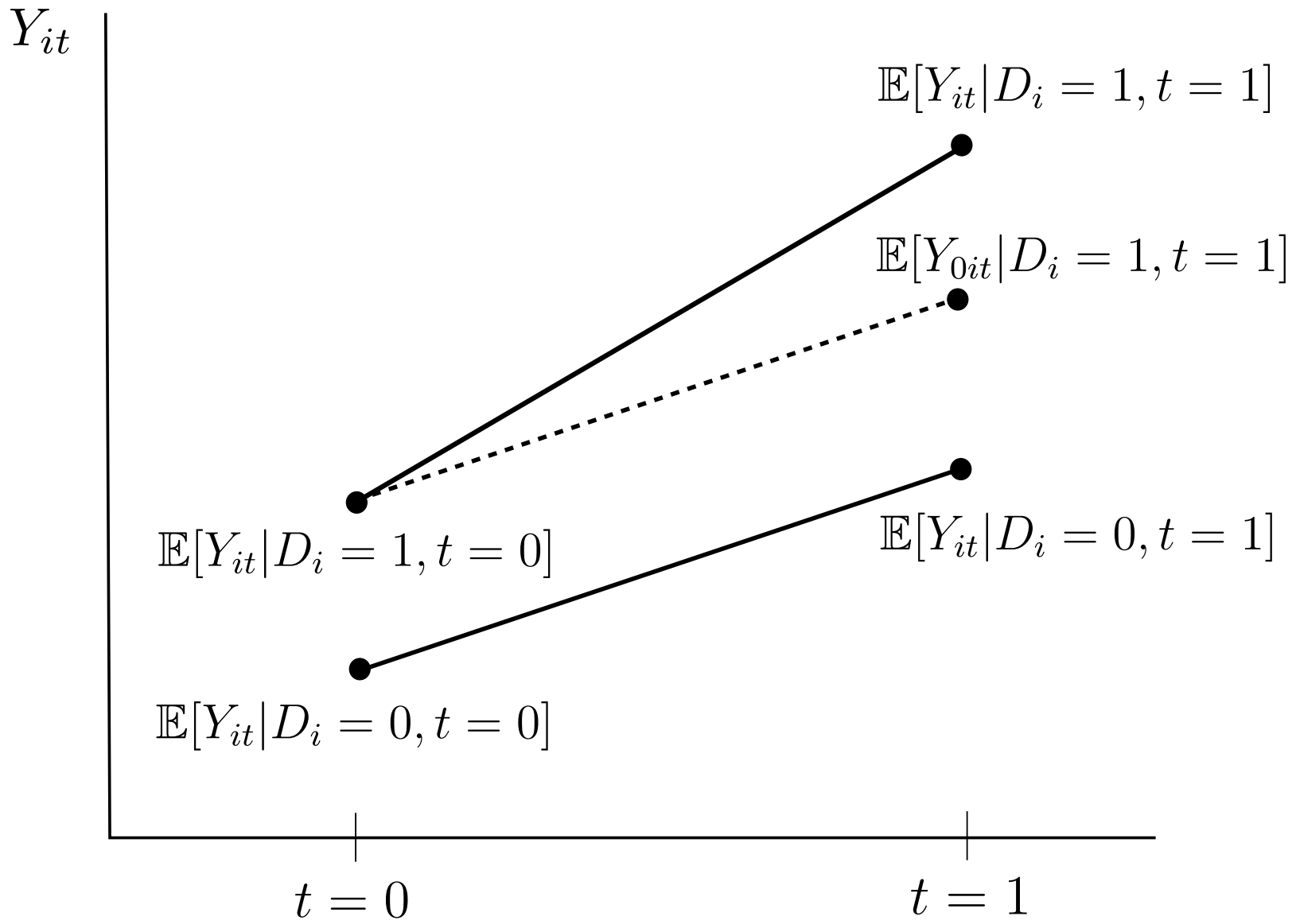
Randomized experiment $\Rightarrow$ simple comparison of the mean outcome in treatment and control groups ("difference" estimator), unbiased and consistent estimate of the ATE.

The approach in this chapter: like in matching or sharp RD, adjusts somehow to **compensate confounders**.

Linking Chapter 2 to treatment effects approaches, we propose an alternative method to eliminate confounders that are **fixed over time** (like a fixed effect), using repeated observations.

Key assumption: common trend.

Difference in differences



Formal discussion

Figure suggests to use **trend observed for untreated** to predict the **counterfactual trend** for treated individuals in the absence of treatment:

$$\mathbb{E}[Y_{0it}|D_i = 1, t = 1] = \underbrace{\mathbb{E}[Y_{it}|D_i = 0, t = 1]}_{\text{level for controls at } t=1} + \underbrace{\{\mathbb{E}[Y_{it}|D_i = 1, t = 0] - \mathbb{E}[Y_{it}|D_i = 0, t = 0]\}}_{\text{difference in levels at } t=0 \text{ difference}}.$$

Fundamental DiD assumption: common trend:

$$\mathbb{E}[Y_{0i1} - Y_{0i0}|D_i = 1] = \mathbb{E}[Y_{0i1} - Y_{0i0}|D_i = 0].$$

Hence, the difference in differences coefficient (which is an average treatment effect on the treated) is:

$$\begin{aligned}\beta = &\{\mathbb{E}[Y_{it}|D_i = 1, t = 1] - \mathbb{E}[Y_{it}|D_i = 1, t = 0]\} \\ &- \{\mathbb{E}[Y_{it}|D_i = 0, t = 1] - \mathbb{E}[Y_{it}|D_i = 0, t = 0]\}.\end{aligned}$$

Diff-in-diff and regression

The difference in differences coefficient can be obtained as the β coefficient in the following **regression**:

$$Y_{it} = \beta_0 + \beta_D D_i + \beta_T T_{it} + \beta D_i T_{it} + U_{it},$$

where $T_{it} = 1$ if individual i is treatment period $t = 1$, and $T_{it} = 0$ otherwise.

With similar arguments as in previous chapters:

- $\beta_0 = \mathbb{E}[Y_{it}|D_i = 0, t = 0]$,
- $\beta_0 + \beta_D = \mathbb{E}[Y_{it}|D_i = 1, t = 0]$,
- $\beta_0 + \beta_T = \mathbb{E}[Y_{it}|D_i = 0, t = 1]$,
- and β is the difference in differences coefficient.

Diff-in-diff and regression

This regression model can be expanded in several ways:

- **Further periods**: In such case, T_{it} is not a time dummy but, instead, a dummy that equals one in the post-treatment period. One could additionally include time effects, but the interaction term should be with the "post" dummy only.
- **Controls**: the regression allows for controls, X_{it} (they work like in regression vs matching).
- **Panel data**: there is actually no need for panel data. However, in the repeated cross-section context, the researcher needs to sustain the assumption that the sample composition does not vary over time, which is satisfied by construction with panel data (also individual fixed effects)
- **Placebo analysis**: a regression that simulates the difference in differences analysis but for a point in time or group of individuals that resemble the treatment period or group but that was actually not treated.

Triple differences

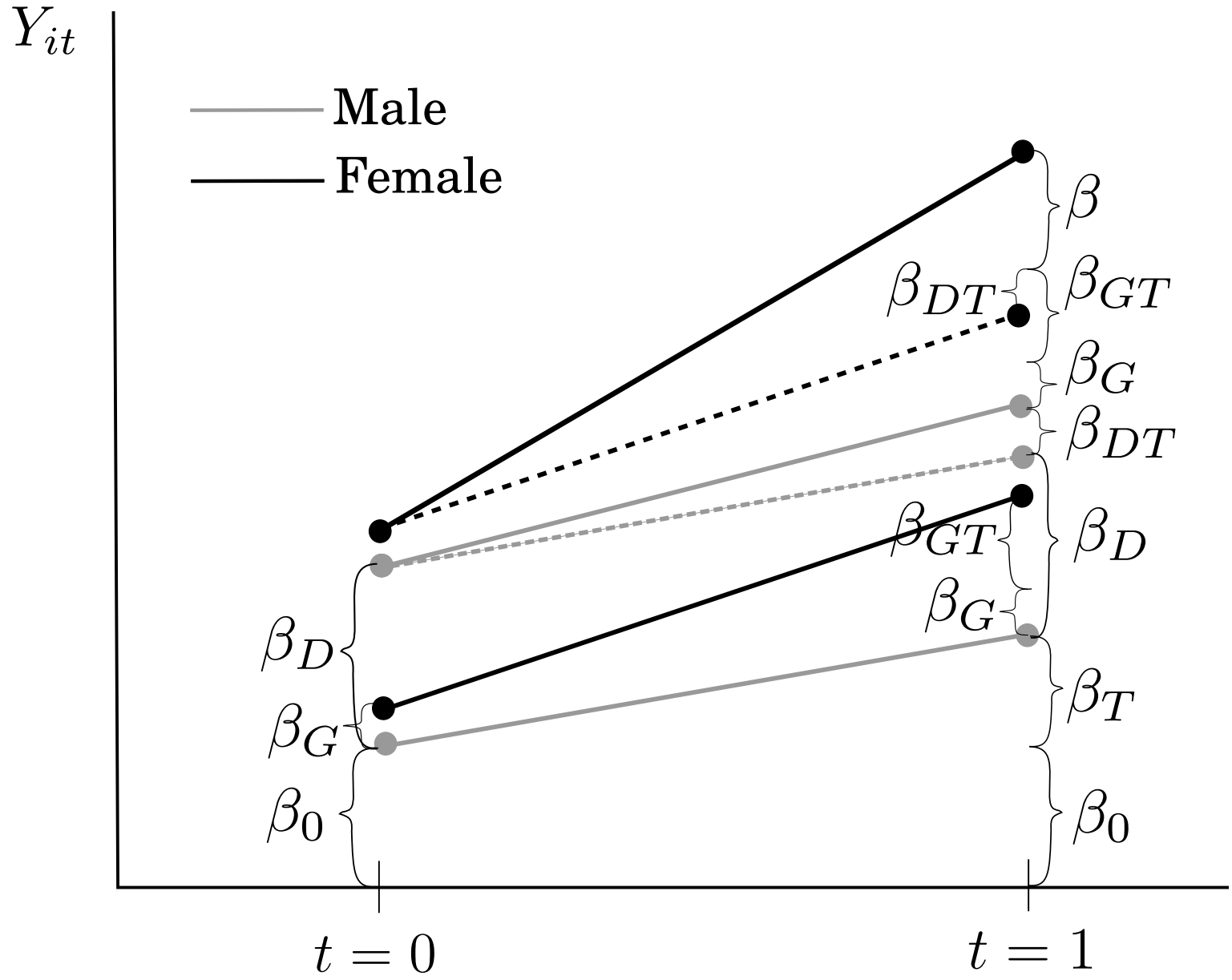
Triple difference: the difference in differences assumption does not hold, but the change in trends is assumed to be the same across sub-groups, some of which should be more affected than others.

Let G_i denote the (say sociodemographic) group to which individual i belongs. Then, the **triple-differences** model is:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_D D_i + \beta_T T_{it} + \beta_G G_i + \beta_{GD} G_i D_i \\ + \beta_{GT} G_i T_{it} + \beta_{DT} D_i T_{it} + \beta G_i D_i T_{it} + U_{it}. \end{aligned}$$

Example: Maternity leave policies combined with a tax reform that affects young and old differently.

Difference in differences



Synthetic Control Methods

Synthetic control methods: use longitudinal data to build the weighted average of non-treated units that best reproduces the characteristics of the treated unit over time prior to the treatment.

Thus, we build an **artificial control** that has the best possible pre-trend possible, and then we compute the difference in differences estimate using such synthetic control group.