Causal Inference in Observational Studies

Contents

Disclaimers: (i) Sections and lines in brown correspond to content which is very much 'under construction'. (ii) For all mathematical simplifications featuring "= ... =", the detailed steps are provided in the Appendix.

1 Definitions

1.1 Research design and identification strategy

Research design = working from the research question, the overall manner in which data will be gathered, assembled and assessed in order to draw conclusions.

In the applied economics literature, and only in the context of (1) *observational* studies that aim to (2) identify a *causal* effect, a subordinate notion is the "identification strategy".[1]

If one's research goal is to identify the causal effect of a specific event or program ("treatment"), and sets internal validity as the priority,[2] then one wants a research design that may credibly identify causal effects.

One such design is a **randomized trial**: an experiment which randomly assigns the participants to either a treatment or a control group. This experiment is often considered as the "gold standard" against which to judge other research designs. In an observational study, one attempts to approximate the force of evidence generated by such an experiment. A key aspect of the research design is hence the identification strategy:

Identification (causal inference) strategy = how *observational* data are used to approximate a real experiment. It is the set of assumptions that will *identify* the causal effect of interest, including:

$$= \begin{cases} \text{— a clear source of identifying variation in a causal variable,} \\ \text{— the use of a particular econometric technique to exploit this information.} \end{cases}$$

1.2 Experiments, natural experiments, quasi-experiments

A true experiment is a study in which the researcher manipulates the level of a treatment (the independent variable of interest) and measures the outcome (the dependent variable of interest). All the important factors that might affect the phenomena of interest are controlled.

A natural experiment is an observational study in which a *randomization* of a treatment D or instrument Z has occurred naturally — mimicking the exogeneity of a randomized experiment. Researchers do not create natural experiments — they find them.

Ex: weather

A quasi experiment is a study of intentional treatment, that resembles a randomized field experiment but lacks full random assignment. Participants are *not* randomly assigned to the treatment or control group. The groups therefore differ in often unobservable ways, so one must control for as many of these differences as possible. The control group is rather called a "comparison" group.

Ex: In the 1990s, the U.S. Department of Housing and Urban Development (HUD) implemented a grant program to encourage resident management of low-income public housing projects. Housing projects were *selected* in 11 cities nationwide, so the treatment (the award of HUD funding) was not randomly assigned. But similar housing projects in the same cities provided a reasonably valid comparison, so the HUD was able to evaluate the program.

[1] Angrist and Pischke (2010) use the notions of research design and identification strategy interchangeably.

[2] True experimental designs may be the "gold standard" of scientific research when considering only internal validity. However, the very methods used to increase internal validity may also limit the generalizability or external validity of the findings. Ex: a zoo is a controllable setting amenable to drawing causal inferences about the behavior of animals, but these inferences may not generalize to the behavior of animals in the wild.

2 Framework: a counterfactual approach to causality

2.1 The potential outcomes framework

2.1.1 The original selection bias problem and the CIA

We have a population, and a treatment of interest $D_i \in \{0,1\}$ whose causal effect we want to estimate. Let Y_i^1, Y_i^0 be individual i's potential outcomes — if they were to receive the treatment or not, respectively — and Y_i their realized outcome. The potential outcomes framework[3] allows us to define quantities:

Individual treatment effects (TE)	$Y_i^1 - Y_i^0, \ \forall i$	*what we'd ideally want to estimate*
Average treatment effect (ATE)	$\mathbb{E}\left[Y_i^1 - Y_i^0\right]$	*what we reasonably want to estimate*
Average treatment effect on the treated (ATET)	$\mathbb{E}\left[Y_i^1 - Y_i^0 \vert D_i{=}1\right]$	*what we reasonably want to estimate*
Difference in average observed outcomes	$\mathbb{E}[Y_i \vert D_i{=}1] - \mathbb{E}[Y_i \vert D_i{=}0]$	*what we can compute*
Each quantity can be made conditional on covariates X; it will be the quantity 'for given X', i.e., within stratum.		

The focus on identification is due to the **original selection bias problem:**

- To measure TE = $Y_i^1 - Y_i^0$, we need to observe the same individual with and without treatment.
- This is impossible, we can never observe the counterfactual.[4] We can only compute the difference in average observed outcomes :

$$\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0] \ = \ ... \ = \ \underbrace{\mathbb{E}\left[Y_i^1 - Y_i^0|D_i{=}1\right]}_{\text{ATET}} \ + \ \underbrace{\mathbb{E}\left[Y_i^0|D_i{=}1\right] - \mathbb{E}\left[Y_i^0|D_i{=}0\right]}_{\text{selection bias}}$$

  The "selection bias" is the average difference in Y_i^0 between the treated and untreated.[5]
- If treatment is randomly assigned, it is independent of potential outcomes: $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i$, so there is no selection bias *in expectation*.[6] The independence assumption (IA) identifies the ATET (and =ATE).
- In observational studies, $(Y_i^0, Y_i^1) \not\perp\!\!\!\perp D_i$. However, if we *match* treated and control individuals to be proper counterfactuals, i.e., if conditional on some pre-treatment characteristics X_i, the assignment of treatment to individuals is independent of potential outcomes: $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i|X_i$, then we can again eliminate selection bias in expectation. We compare outcomes within each stratum of X_i:

$$\underbrace{\mathbb{E}[Y_i|D_i{=}1, X_i] - \mathbb{E}[Y_i|D_i{=}0, X_i]}_{\text{diff. in average outcomes for given } X_i} \ = \ ... \ = \ \underbrace{\mathbb{E}\left[Y_i^1 - Y_i^0|D_i{=}1, X_i\right]}_{\text{ATET for given } X_i} \ + \ \underbrace{\mathbb{E}\left[Y_i^0|D_i{=}1, X_i\right] - \mathbb{E}\left[Y_i^0|D_i{=}0, X_i\right]}_{\text{selection bias for given } X_i}$$

  The conditional independence assumption (CIA) eliminates the last term, and so identifies the ATET for each value of X_i, i.e., within each subpopulation. We can then combine these ATETs by weighting them in our preferred way to recover a single ATET.[7]

[3] The potential outcomes framework for causal inference builds on (Neyman, 1923), was extended to observational studies by (Rubin, 1974), and became popular in econometrics around 1990. One strong assumption is that of no interference between units: the TE on one unit is independent of the treatment received by others. This excludes spillovers, strategic interactions...

[4] This is the 'fundamental problem of causal inference'. Its implication: we *never* observe causal effects.

[5] For example: if individuals with low Y_i^0 choose treatment more frequently, then $\mathbb{E}\left[Y_i^0|D_i{=}1\right] < \mathbb{E}\left[Y_i^0|D_i{=}0\right]$. Comparing Y between treated and untreated underestimates the TE. Say we look at the effect of hospitalization; sick individuals go to the hospital (get treated) more often than healthy individuals. But they would also have been less healthy had they stayed at home.

[6] Independence removes selection bias *in this expectation form*, where the expectation is taken over repeated randomizations on the trial sample, each with its own allocation of treatments and controls (Deaton and Cartwright, 2018). Independence does not imply actual balance in any single trial: the sample analog of the last term (i.e., the net differences of means of all the other causes across the two groups) may not be 0.

[7] For instance, the matching estimand will weight them by the distribution of covariates among the treated, whereas the linear regression estimand will weight them by the variance of treatment — see section 2.1.2.

We can estimate an **unbiased** causal effect iff an **identifying/independence**[8] **assumption** holds:

— if IA $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i$ $\implies$ we can estimate the ATET.

— if ~~IA~~ but CIA $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i | X_i$ $\implies$ we can estimate the ATET in each stratum.

— if ~~CIA~~ but $\exists$ a relevant instrument Z that is an exogenous source of variation in D:

$(Y_i^0, Y_i^1) \perp\!\!\!\perp Z_i, \quad Z_i \not\perp\!\!\!\perp D_i$ $\implies$ we can estimate *a* LATE.

So we need an **identification strategy** that convinces us that an IA holds.

Note: The identification result extends beyond average treatment effects. Independence means that the entire distribution of potential outcomes is independent of the treatment, s.t. we can also estimate quantile treatment effects — i.e, $\forall p \in [0,1]$, the effect of the treatment at quantile p: $\tau_p \equiv \mathbb{Q}_{Y^1}(p) - \mathbb{Q}_{Y^0}(p)$.[9] Quantile treatment effects may be informative if TEs are concentrated in tails of the distribution of outcomes, or provide more robust estimates than ATEs in settings with thick-tailed distributions.

2.1.2 Expressing TE as a linear regression

Suppose a heterogeneous TE, i.e., $Y_i^1 - Y_i^0 = \beta_i$. Note β the average for the treated population $\mathbb{E}[\beta_i | D_i{=}1]$, i.e., the ATET. The relation between observed outcomes and potential outcomes (how we estimate our TE) can be written as a linear regression on the treatment:

$$\begin{aligned} Y_i &= Y_i^0 + \left(Y_i^1 - Y_i^0\right) D_i \\ &= Y_i^0 + \quad \beta_i D_i \\ &= Y_i^0 + (\beta_i - \beta + \beta) D_i + \mathbb{E}[Y_i^0] - \mathbb{E}[Y_i^0] \\ &= \mathbb{E}[Y_i^0] + \beta D_i + (\beta_i - \beta) D_i + Y_i^0 - \mathbb{E}[Y_i^0] \\ &= \quad \alpha \quad + \beta D_i + \qquad\qquad u_i \end{aligned}$$

We can show that the OLS slope estimand $\beta_{\text{OLS}} \equiv \frac{\text{cov}[Y_i, D_i]}{\mathbb{V}[D_i]}$ simplifies to $\mathbb{E}[Y_i | D_i{=}1] - \mathbb{E}[Y_i | D_i{=}0]$: the difference in average observed outcomes. This, in turn, given the regression equation, equals $\beta + \mathbb{E}[u_i | D_i{=}1] - \mathbb{E}[u_i | D_i{=}0] = ... = \beta +$ selection bias.

β_{OLS} recovers the ATET iff there is no selection bias, or equivalently, iff u_i is uncorrelated with D_i.
An identification problem (dependence) $\iff$ a regression problem (endogeneity).

Is the linear regression always appropriate? The demonstration above corresponds to the simplest setting: an unlimited Y, a binary D, and no X. Does β_{OLS} still recover an ATE when we depart from that?

- **With covariates X** We want to estimate a difference in averages. The more covariates there are, the more a nonparametric analysis is complicated, so we turn to regression as a computational device. However, there are differences between the matching and the regression estimands: they sum the estimates of the within-stratum ATETs $\delta_x \equiv \mathbb{E}[Y_i | D_i{=}1, X_i] - \mathbb{E}[Y_i | D_i{=}0, X_i]$ into an overall ATET using different weights (Angrist and Pischke, 2008, 3.3.1). This will result in two different estimates only if δ_x varies along X, i.e., if the TE is heterogeneous.[10] For simplicity, consider a discrete X_i:

[8] Independence assumptions are also called "ignorability" assumptions in statistics, meaning ignorability of the assignment mechanism. Indeed, with independence, we don't need to model the assignment process to estimate causal effects, we need only compare group means. Examples of assignment mechanism: random assignment ($\Leftrightarrow$ IA); selection on observables ($\Leftrightarrow$ CIA); selection on unobservables...

[9] ⚠ The p-th QTE $\mathbb{Q}_{Y^1}(p) - \mathbb{Q}_{Y^0}(p)$ is the effect of the treatment at quantile p, i.e., a difference between quantiles of the two marginal potential outcome distributions. *Not* the p–th quantile of the treatment effect $\mathbb{Q}_{Y^0 - Y^1}(p)$. In general, the latter quantile of the difference differs from the difference in the quantiles.

[10] Then we may want to interact D with X! See section 4.3.

– In the matching estimand, the weights are proportional to the conditional probability of treatment:

$$\beta_{\text{M}} = ... = \frac{\sum_x \delta_x w_{\text{M}}}{\sum_x w_{\text{M}}}, \quad w_{\text{M}} \equiv \text{P}(D_i{=}1|X_i{=}x)\ \text{P}(X_i{=}x)$$

– In OLS, they are proportional to the conditional variance of treatment — which is maximized when $P(D_i{=}1|X_i) = .5$, i.e., for values of X_i with as many treated as control observations. OLS gives more weight to more precise within-strata estimates:

$$\beta_{\text{R}} = ... = \frac{\sum_x \delta_x w_{\text{R}}}{\sum_x w_{\text{R}}}, \quad w_{\text{R}} \equiv \text{P}(D_i{=}1|X_i{=}x)\ \big(1 - \text{P}(D_i{=}1|X_i{=}x)\big)\ \text{P}(X_i{=}x)$$

- **With multiple treatment intensities** We can generalize the CIA to settings where the treatment has more than two levels, and still use linear regression to recover causal effects. Note $\{d\}$ the set of possible treatment intensities, $Y_i^d \equiv f_i(d)$ the potential outcome of individual i under exposure to level d, and d_i and $Y_i \equiv f_i(d_i)$ its realized treatment intensity and outcome. The CIA becomes $Y_i^d \perp\!\!\!\perp d_i|X_i$, and the within-stratum ATET for a 1-unit increase in d is $\delta_x \equiv \mathbb{E}[f_i(d) - f_i(d{-}1) \mid d_i{=}d, X_i{=}x]$.

  – Simplest case: If $f_i()$ is linear in d and is the same for everyone (i.e., parameters aren't indexed by i), up to an error: $f_i(d) = \alpha + \beta.d + X_i'\gamma + e_i$, then the regression model $Y_i = \alpha + \beta.d_i + X_i'\gamma + e_i$ estimates the ATET.

  – Generalization: If $f_i()$ isn't linear in d and/or differs across people, regression still estimates an average causal effect: the weighted average of the individual-specific difference $f_i(d) - f_i(d{-}1)$.

- **With limited *Y*** (e.g., binary, positive) Regardless of whether Y is limited, in Angrist and Pischke (2008)'s view, **linear regression** is always legitimate as it **provides the best (MMSE) linear approximation to the conditional expectation function (CEF)**.[11]

  – Simplest case (binary D, no X):
  $\mathbb{E}[Y_i|D_i]$ is inherently a linear function of D, so the regression vector $D_i'\beta_{\text{OLS}}$ and the CEF are exactly equal, regardless of the constraints on Y. β_{OLS} of the linear regression of Y on D is $\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0]$ — which identifies the ATET under the IA.

  – Generalization (D isn't a dummy and the CEF includes other covariates):
  $\mathbb{E}[Y_i|D_i, X_i]$ is generally nonlinear for limited Ys (Y isn't normal so (Y, D, X) isn't multivariate normal, and we rarely have enough data to run the saturated-covariate specification). So linear regression won't fit the CEF perfectly. But it still provides the MMSE approximation to the CEF, and the latter is causal under the CIA! It will miss some features of the CEF and so may generate fitted values outside Y's boundaries,[12] but when it comes to marginal effects (the average changes in CEF) this might matter little.

2.1.3 Why might the IA/CIA not hold? Endogeneity

In the simple (linear & univariate) regression model $y_i = \alpha + \beta x_i + e_i$, the variable x_i is

- **endogenous** if it is correlated with the error term: $\text{cov}[x_i, e_i] \neq 0$.

[11] A good summary of the empirical relationship between Y and D is the CEF $\mathbb{E}[Y_i|D_i]$, and Least Squares regression appproximates the CEF. Recall the OLS problem: $\hat{\beta}_{\text{OLS}} \equiv \text{argmin} \sum_{i=1}^n (Y_i - D_i'\beta)^2$.
– If the CEF is linear, i.e., $\mathbb{E}[Y_i|D_i] = D_i'\beta$, then the regression function $D_i'\beta_{\text{OLS}}$ is the CEF: $\beta_{\text{OLS}} = ... = \beta$. A CEF is linear under only 2 rare circumstances:
  * if (Y_i, D_i) is multivariate Normal, or
  * if the model is saturated (i.e., it has a separate parameter for each possible combination of values of the regressors D_i).
– The slope vector $D_i'\beta_{\text{OLS}}$ provides the best — meaning minimizing mean squared errors — *linear* approximation to the CEF. Linear regression is therefore always a useful descriptor of a CEF. Though one could object that if the CEF is highly nonlinear, what would we really learn from a linear approximation to it?

[12] A well-known problem of the linear probability model, whereas nonlinear models like Probit and Tobit produce CEFs that respect the $[0, 1]$ boundaries.

- **exogenous** if it is uncorrelated with the error term: $\text{cov}[x_i, e_i] = 0$.

If x is endogenous, the OLS slope estimator of β will comprise not only the partial derivative w.r.t. x (what we want) but also an indirect effect through e: $\beta_{\text{OLS}} = \frac{dy(x,e)}{dx} = \frac{\partial y}{\partial x} + \frac{\partial y}{\partial e}\frac{\partial e}{\partial x} = \beta + \frac{\partial e}{\partial x} \neq \beta$. The OLS estimator is therefore biased and inconsistent for β.

In our case of interest, if the treatment D_i is endogenous, i.e., $\text{cov}[D_i, e_i] \neq 0$, it means there is an imbalance in potential outcomes across the treatment groups. The CIA doesn't hold. Our estimate will be biased.

Sources of endogeneity

- reverse causality or simultaneity: If y also affects D, this is captured by e, making e correlated with D;
- non-random measurement error in D — specifically, that is correlated with y;
- omitted variable bias (OVB): All omitted variables[13] are captured by e. Therefore, if an omitted variable w is correlated with D, e is correlated with D. w is a "confounding variable".
  This source of endogeneity is the most common, and therefore the one we will focus on in the rest of the document.

→ In observational studies,
– excluding a confounding variable creates bias, so we must adjust for all *confounders*.
– with all confounders assumed to be adjusted for, we estimate an unbiased causal effect.
– because we can rarely be certain to have accounted for all confounders,[14] we turn to alternative causal inference or **"identification" strategies**, that rely on other assumptions.

[13] An omitted variable is an explanatory variable not included in the regression but which is a determinant of y.

[14] For instance, in cross-sectional approaches, we worry about time-invariant omitted variables. As a cross-section offers only 'across' (inter-individual) variation, if Y is affected by unobservable variables that systematically vary across groups, our estimate will be biased. With panel data, we have across and 'within' (intra-individual) variation. Using individual fixed effects, we can focus on within variation, which greatly reduces the threat of OVB.

2.2 The causal graph framework

Pearl (2009) proposes an alternative to the potential outcomes framework for thinking about causality: a causal graph framework.[15] We introduce it here on the basis of two important points:

1. The two frameworks are not opposed, they both define causality using counterfactuals. A causal effect is a comparison between two states of the world: a realized state as the intervention took one value, and a "counterfactual" state that would have happened had the intervention taken another value.
2. Each encodes these counterfactual causal states differently, DAGs are just a language to make the identifying assumptions explicit. Add Imbens (2020).

The potential outcome and the causal graph frameworks are therefore complementary perspectives, and it can be useful to frame one's causal inference in the language of each framework.[16]

2.2.1 Elements of directed acyclic graphs (DAGs)

Relationships are encoded with nodes and edges. In the context of analyzing the causal effect of a treatment variable D on an outcome variable Y, we introduce additional notions:

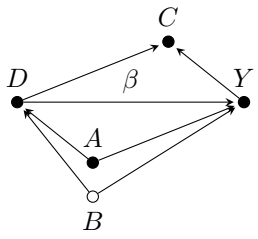
- a node represents a random variable; a solid circle if it is observed, hollow otherwise;
- all edges are directed and represent causal relationships;
- a path is any sequence of edges;
- a **back-door path** = any path between D and Y that begins with an arrow pointing to D.[a] It is *closed* if at least one variable along the path is observed, *open* otherwise.
- a **confounder** = a variable that determines both D and Y along some path.
  $\Rightarrow$ *Fluctuations in the confounder drive some of the association between D and Y; the total association between D and Y is therefore not equal to β.*
- a **collider** = a variable that is determined by both D and Y along some path.
  Colliders do not generate an unconditional association between D and Y, i.e., bias, so one need not adjust for them. On the contrary, including them would generate bias.

[a] This path is "entering D through the back door".

Importantly, a DAG is a *complete* encoding of assumptions about causal relationships: those assumed to exist represented by arrows, and those assumed to not exist represented by missing arrows. I.e., the exclusion of an arrow is not the absence of an assumption, but the assumption that there is no direct relationship.

For example, the basic DAG on the right encodes:

* explicitly, 4 paths linking D to Y:
  $D \xrightarrow{\beta} Y$: a direct (causal) path
  $D \longleftarrow A \longrightarrow Y$: a back-door confounding path, closed
  $D \dashleftarrow B \dashrightarrow Y$: a back-door confounding path, open
  $D \longrightarrow C \longleftarrow Y$: a colliding path
* implicitly, 3 assumptions of no direct relationships between A, B and C.



[15] For a detailed presentation, see Morgan and Winship (2015, ch. 1.5 & 3), of which this section is (an attempt of) a summary.

[16] How directed graphs encode (potentially counterfactual) causal states is not detailed here. See sections 3.4 and 3.6 of Morgan and Winship (2015), or Pearl (2009), for a detailed presentation. Importantly, we also consider only the subset of directed *acyclic* graphs (DAGs), where no directed paths emanating from a causal variable also terminate at the same causal variable. This prohibition of cycles notably rules out representations of simultaneous causality and feedback loops. Section 3.2 of Morgan and Winship (2015) discusses the implications.

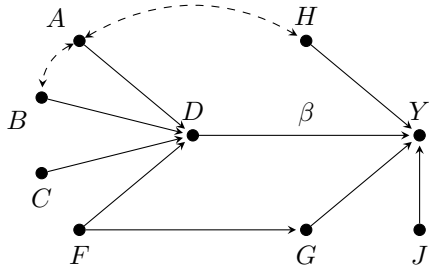
2.2.2 Two identification strategies: 1. blocking back-door paths; 2. instruments

We want to estimate the causal effect of a treatment D on Y. We represent in a DAG this causal relationship, and all other relationships relevant to the effect of D on Y. *Given the structure of the causal relationships, which variables must we observe and include to estimate the causal effect of D on Y?*

- Strategy 1: blocking back-door paths
  The most common concern with observational data is that D and Y are partly determined by a third variable, i.e., that there is a back-door path. **The total association between D and Y equals β iff there are no back-door paths.**
  - In the previous basic DAG:
    - Assume B wasn't there. The only back-door path between D and Y is closed as we observe A. If we adjust for A, i.e., hold it fixed, we remove the association between D and Y that is driven solely by fluctuations in A, and recover the causal effect β. **We can recover β by *blocking* all back-door paths**, i.e., conditioning on one confounder along each back-door path.
    - However, the back-door path through B is *open*, as B is unobserved. → We cannot recover β by blocking back-door paths.
  - In the more complex DAG on the right, there are three back-door paths:[17]

    $D \longleftarrow A \dashleftarrow\dashrightarrow H \longrightarrow Y$
    $D \longleftarrow B \dashleftarrow\dashrightarrow A \dashleftarrow\dashrightarrow H \longrightarrow Y$
    $D \longleftarrow F \longleftrightarrow G \longrightarrow Y$

    We can block all back-door paths by either:
    - conditioning on H and either F or G
    - conditioning on A and B,[18] and either F or G



- Strategy 2: instruments
  Instead of blocking back-door paths to estimate β directly, we can leverage an exogenous shock to D to estimate β indirectly. We use exogenous variation in an instrument Z[19] to isolate covariation in D and Y. In the DAG above, we can use as instrument for D either C, or F after conditioning on G.

To estimate the effect β of D on Y, we reach the same conclusion as with the potential outcomes framework:

> → In observational studies,
> – leaving a back-door unblocked, i.e., excluding a confounding variable, creates bias, so we must block all back-doors (adjust for all confounders).
> – *"Back-door criterion"*: with all back-doors blocked, i.e., all confounders conditioned on, we estimate an unbiased causal effect.
> – because we can rarely be certain that we have accounted for all confounders, we turn to alternative causal inference or **"identification" strategies**, that rely on other assumptions.

[17] To show that two variables are mutually dependent on one or more unobserved common causes, instead of abiding by the definitions and showing it with U as in the left figure below, we can use a curved dashed bidirected edge as in the right figure as a shorthand. These bidirected edges should however not be interpreted as mere correlations between the two variables, they represent an unspecified set of unobserved common causes of the two variables that they connect.

[18] Conditioning only on A would not suffice. As A is a collider along the path between B and H, conditioning only on A would create dependence between B and H, and so wouldn't eliminate the noncausal association between D and Y.

[19] Instruments are formally introduced in the next section. In short, a variable Z is a valid instrument for D if it causes D but does not have an effect on Y except through its effect on D. We can then estimate consistently the effect of D on Y by taking the ratio of the relationships $Z \leftrightarrow Y$ and $Z \leftrightarrow D$.

Advantages of DAGs

- DAGs are helpful at clarifying the relationships between variables and guiding a research design that has a shot at identifying a causal effect. They force us to write all our assumptions, notably all the relationships that we assume are null between variables of importance. A DAG is telling two stories: what is happening, and *what (we assume) is not happening.*
- DAGs encode causal relationships that are completely nonparametric. When considering analysis strategies, it is thereby not necessary to make assumptions about the functional form of the dependence of Y on the variables that cause it. This notably means that all interactions between the effects of different variables on Y are implicitly permitted. No new arrows are needed to represent these interactions — where, for example, the effect of D on Y varies with the level of X — as the directed edges only signify inclusion in the structural function $f_Y(D, X, ...)$.
- DAGs show that there is often more than one way to estimate a causal effect, and that "controlling for all other causes of Y" can be misleading. In DAG #2, there were two completely different and relevant conditioning strategies (after conditioning for either F or G): conditioning either on H or on A and B. They also show clearly the importance of collider variables: endogenous variables that must be handled carefully — or they may create conditional dependence that can sabotage a causal analysis.
- They are helpful for communicating research designs; pictures do speak a thousand words.
- They provide a bridge between empirical schools, such as structural and reduced form.

2.3 Learning the causal structure vs the magnitude of effects given the structure

Gelman (2011) states that in the social sciences, from a given identification strategy, one cannot reliably learn the causal structure of relationships but only these relationships' magnitudes given the model.

Add `https://theeffectbook.net/ch-EventStudies.html#the-joint-test-problem`

3 Design stage: Applied identification methods

Hierarchy of common identification methods A contestable hierarchy of the most common identification methods in the 'randomista' toolkit, based on their capacity to mimic random assignment, is as follows:

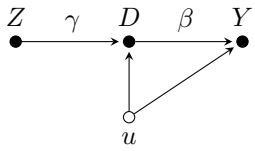
0. Randomized experiment (RCT) — or 'natural' randomization of treatment D
1. Instrumental Variables (IV) and regression discontinuity (RD)
   If there may be selection into treatment based on unobservables, we use an instrument or discontinuity that induces quasi-experimental variation in treatment status.
2. Difference-in-differences (DiD) and event studies
   If we have repeated observations and want to estimate the effect of an event, we use research designs that assume parallel trends and the presence of only time-invariant confounders.
3. Matching estimators
   Strategies based solely on matching are considered much less credible — in terms of making us believe in the CIA, and thus their ability to recover a causal effect — than strategies based on some exogenous variation. However, matching is a type of procedure that can complement a natural-/quasi-experiment design. It is adressed in section 4.

The sections below present, for each method, in the canonical setup: (i) the assumed data generating process (DGP), (ii) the identifying assumptions, (iii) the estimand, i.e., the treatment effect of interest, (iv) the estimator used, and (v) some best practices, and strengths and weaknesses. Importantly, the **relation between the actual observed outcomes Y_i and the conceptual potential outcomes Y_i^0, Y_i^1** is made explicit. This relation is crucial: it represents how our estimation (using Y_i) is able to recover a causal treatment effect (defined with Y_i^0, Y_i^1).

For simplification purposes, all methods are presented without the inclusion of exogenous controls X_i, but the relationships can be generalized to conditioning on covariates X_i.

3.1 IV

Data Generating Process (DGP) $Y_i = \alpha + \beta_i\, D_i + u_i, \quad \text{cov}[D_i, u_i] \neq 0$: D_i is endogenous. But $\exists$ a binary instrument Z_i that is a random source of variation in D_i, it "assigns treatment" or changes the probability of treatment. We can use the instrument to isolate variation in D which is unrelated to u, and recover β.[20]



$$D_i = \delta + \gamma\ Z_i + v_i$$
$$Y_i = \alpha + \beta\ D_i + u_i, \qquad \text{cov}[D_i, u_i] \neq 0$$

Potential outcomes: We define the treatment assignment $Z_i \in \{0,1\}$ and the treatment realization $D_i \in \{0,1\}$. $Z_i = 0$ induces the potential treatment status D_i^0, realized as 0 if individuals comply, 1 if not. $Z_i = 1$ induces D_i^1, realized as 1 if they comply, 0 if not. The compliance behavior defines 4 categories of participants — which the researcher *cannot* observe; they can only observe the assignment Z_i and the realization D_i.

	D_i^0	D_i^1
compliers	0	1
always-takers	1	1
never-takers	0	0
defiers	1	0

Identifying assumptions

(A1) independence w.r.t. the potential outcomes, i.e., $cov[Z_i, v_i] = 0$

(A2) exclusion restriction: $\text{cov}[Z_i, u_i] = 0$ *(and* $cov[Z_i, covariates_i] = 0$*)*. I.e., Z affects Y only through D.

(A3) relevance: $\text{cov}[Z_i, D_i] \neq 0$

(A4) monotonicity: the instrument does not discourage treatment (no defiers). This assumption is weaker (and therefore more realistic) than the assumption of homogeneous effects.

Estimand $$\beta_{\text{IV}} \equiv \frac{\text{cov}[Y_i, Z_i]}{\text{cov}[D_i, Z_i]} = ... = \frac{\mathbb{E}[Y_i \mid Z_i{=}1] - \mathbb{E}[Y_i \mid Z_i{=}0]}{\mathbb{E}[D_i \mid Z_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0]} = ... = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid D_i^0{=}0, D_i^1{=}1]}_{\textbf{LATE on the compliers}}$$

Estimator A natural choice of estimator is the sample analog called "Wald estimator" $\hat{\beta}_{\text{W}} = \frac{\widehat{\text{cov}}[Y_i, Z_i]}{\widehat{\text{cov}}[D_i, Z_i]}$. Note that:

– The slope estimate $\widehat{\gamma}_{\text{LS}} = \frac{\widehat{\text{cov}}[D_i, Z_i]}{\widehat{\mathbb{V}}[Z_i]}$ from regressing D on Z consistently estimates $\gamma = \frac{\text{cov}[D_i, Z_i]}{V[Z_i]}$

– The slope estimate $\widehat{\gamma.\beta}_{\text{LS}} = \frac{\widehat{\text{cov}}[Y_i, Z_i]}{\widehat{\mathbb{V}}[Z_i]}$ from regressing Y on Z consistently estimates $\gamma.\beta = \frac{\text{cov}[Y_i, Z_i]}{V[Z_i]}$

$\implies$ The ratio $\frac{\widehat{\gamma.\beta}_{\text{LS}}}{\widehat{\gamma}_{\text{LS}}} = \frac{\widehat{\text{cov}}[Y_i, Z_i]}{\widehat{\text{cov}}[D_i, Z_i]} = ... = \beta + \frac{\widehat{\text{cov}}[u_i, Z_i]}{\widehat{\text{cov}}[D_i, Z_i]} \xrightarrow[n \to +\infty]{p} \beta$: is consistent but biased.

$\hat{\beta}_{\text{W}}$ turns out to be numerically equivalent to the two-stage least squares (2SLS) estimator $\hat{\beta}_{\text{2SLS}}$ obtained through the two-step process:[21]

$$\text{1st stage:} \quad D_i = \delta + \gamma.Z_i + v_i \implies \widehat{D}_i = \widehat{\mathbb{E}}[D_i | Z_i]$$
$$\text{2nd stage:} \quad Y_i = \widetilde{\alpha} + \widetilde{\beta}.\widehat{D}_i + e_i$$

Note: The reduced form of a model is the one in which the endogenous variables are expressed as functions of the exogenous variables. In the IV setting, the regression of y on Z is therefore called the reduced form. Its estimates correspond to the intention to treat (ITT), whereas the IV estimates the treatment on the treated.

[20] For more complicated treatment variables, we will need more complicated instruments. To identify *several* treatment variables, we will need at least as many instruments. To identify a *continuous* treatment, we can't use a binary instrument.

[21] The point estimates are equivalent, however the SEs of the 2nd stage would not give the correct SEs, as we need to adjust for the two stages of estimation. We must account for the estimation uncertainty from the first-stage (the first-stage is based on a sample, not the population, making $\widehat{D_i}$ a random variable, instead of the usual fixed variable). Most 2SLS packages do the adjustment automatically — otherwise one can simply bootstrap the SEs manually.

Best practices

- Support the relevance assumption by showing a large F-statistic for the 1st stage (rule of thumb: $F > 10$). The bigger F, the "stronger" the instrument.
- As in any observational study, adjust for all other relevant pre-treatment variables, making sure to include the same variables in both stages.
- Different valid instruments select different sets of compliers, leading to different estimands and thus estimates. Think of the group of compliers selected, to make sure the instrument is relevant w.r.t. the policy of interest. Then count and characterize these compliers to get more out of the LATE.
- For models that are non-linear in D, the properties of 2SLS do not necessarily hold, so one may want to consider alternative estimation strategies. Ex: the "control function method" (2 stages: (i) same first stage, extract the residuals $\hat{v}$; (ii) regress y on $(Z, D, \hat{v})$, estimate by OLS). Limits: CF is generally more efficient but less robust than 2SLS as it imposes additional restrictions.

Strengths & weaknesses

+ Compelling identification strategy
– Strong assumptions
– $\hat{\beta}_{\text{IV}}$ is less efficient than OLS, and this precision further decreases with weak instruments.
– $\hat{\beta}_{\text{IV}}$ has "finite sample bias", which increases with the weakness and number of instruments.
$\implies$ Beware of weak instruments. They can render $\hat{\beta}_{\text{IV}}$ considerably less efficient and even more biased than $\hat{\beta}_{\text{OLS}}$.[22] See Andrews et al. (2019).

Counting and characterizing compliers to get more out of the LATE Compliers ($D_i^1 > D_i^0$) are rarely representative of the population, due to selective uptake. While we cannot identify individual compliers in the data, we can estimate the size of the complier group, and characterize them in terms of their distribution of observed covariates.

- Counting compliers: We can measure (Angrist and Pischke, 2008, 4.4.4):
  - The size of the complier group. It is the Wald 1st stage: $P[D_i^1 > D_i^0] = ... = \mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0]$
  - The share of treated that are compliers:

$$P[D_i^1 > D_i^0 \mid D_i{=}1] = \ldots = \frac{P[Z_i{=}1] \times (\mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0])}{P[D_i{=}1]} = \frac{\text{share}(Z_i{=}1) \times \text{1st stage}}{\text{share treated}}$$

- Characterizing compliers: We can describe the distribution of covariates X for compliers.
  - For binary characteristics, we can calculate relative likelihoods (Angrist and Pischke, 2008, 4.4.4). For example, the likelihood that a complier has $X_i = 1$ relative to any individual is:

$$\frac{P[X_i{=}1 \mid D_i^1 > D_i^0]}{P[X_i{=}1]} = \ldots = \frac{\mathbb{E}[D_i|Z_i{=}1, X_i{=}1] - \mathbb{E}[D_i|Z_i{=}0, X_i{=}1]}{\mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0]} = \frac{\text{1st stage} \mid X_i{=}1}{\text{1st stage}}$$

  - For general covariates, we can calculate the mean —or other features of the distribution — of the covariate for compliers using Abadie (2003)'s kappa-weighting scheme:
    Suppose the identifying assumptions hold conditional on X_i. For any function $g(Y_i, D_i, X_i)$ with finite expectation, we have $\mathbb{E}[g(Y_i, D_i, X_i) \mid D_i^1 > D_i^0] = \frac{\mathbb{E}[\kappa_i g(Y_i, D_i, X_i)]}{\mathbb{E}[\kappa_i]}$, with the weighting function $\kappa_i = 1 - \frac{D_i(1-Z_i)}{1-P[Z_i=1|X_i]} - \frac{(1-D_i)Z_i}{P[Z_i=1|X_i]}$

Application: Kowalski (2021).

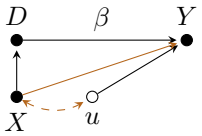
[22] Note that these are different biases: endogeneity bias in the case of the OLS estimator and sample bias with the IV.

3.2 RD

Known assignment mechanism but no overlap

Sharp RD

DGP Treatment D_i is not randomly assigned, it is deterministic, but *discontinuous* along a continuous "running variable" X_i, s.t. there is "local randomization" around a cutoff c: $D_i = \mathbb{1}\{X_i \geq c\}$. Because D_i is a deterministic function of X_i, there are no confounding variables other than X_i. Given the trend relation $\mathbb{E}[Y_i^0 \mid X_i] = f(X_i)$, the DGP is described below, where the brown arrows disappear as $X \to c$:[23]



$$Y_i = \alpha + \beta.D_i + f(X_i, \phi) + u_i$$

⚠ There is zero overlap (no value of X_i with both treatment and control observations), so we must extrapolate across X_i. This means the RD estimate will be only as good as our model for $\mathbb{E}[Y_i^0|X_i]$: we can't be that agnostic about functional form. By looking only at data in a small neighborhood around c, the TE estimate should not depend much on the correct specification of that model.

Identifying assumptions

(A1) *local* continuity: the expected potential outcomes $\mathbb{E}[Y_i^1|X_i]$ and $\mathbb{E}[Y_i^0|X_i]$ are continuous in X_i at c. I.e., the other determinants of Y don't jump at c. $\implies$ The average outcome of those right below the cutoff (who are denied the treatment) are a valid counterfactual for those right above (who receive it).

(A2) relevance: discontinuity in the dependence of D_i on X_i: $D_i = \mathbb{1}\{X_i \geq c\}$

I.e., if there appears to be no other reason for Y_i to be a discontinuous function of X_i, we can attribute a jump in Y_i at c to the causal effect of D_i.

Estimand $\beta_{\text{RD}} = \lim_{x \to c^+} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \to c^-} \mathbb{E}[Y_i \mid X_i = x] = \ldots = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = c]}_{\textbf{LATE at the cutoff}}$

Estimator We can estimate β at the cutoff by running the centered regression below:[24]

$$Y_i = \alpha + \boldsymbol{\beta}\, D_i + f(X_i - c) + e_i$$

Best practices

- Choice of $f()$: $f()$ is unknown. This is a problem, as misspecification of the functional form of the DGP may bias the estimate. Estimation is therefore done with flexible functional forms, such as:
  - a local linear regression model: $Y_i = \alpha + \boldsymbol{\beta} D_i + \gamma_1(X - c) + \gamma_2(X - c)D + e_i$ with $c - h \leq X \leq c + h$.[25]
  - a polynomial regression model with a low-degree polynomial, e.g., quadratic. Higher-order polynomials can lead to overfitting and introduce bias (Gelman and Imbens, 2019).

  In both cases, report the results of several specifications to assess the sensitivity to $f()$.
- As in any observational study, adjust for all other relevant pre-treatment variables. Just because the treatment assignment depends on X, there is no reason to expect overlap and balance across other pre-treatment characteristics. We need to adjust for pre-treatment differences between the two groups.

[23] The causal graph is taken from Steiner et al. (2017).

[24] To allow for different trend functions for $\mathbb{E}[Y_i^0|X_i]$ and $\mathbb{E}[Y_i^1|X_i]$ (i.e., to let the regression model differ on each side of the cutoff), add interactions between D and $f()$: $Y_i = \alpha + \beta.D_i + f(X_i, \phi_l) + f(X_i, \phi_r)D_i + e_i$

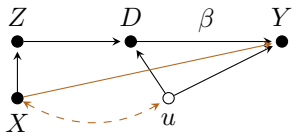
[25] A larger bandwidth h increases precision but also bias. Choose the optimal h by estimating the model's predictive accuracy for different values of h, for example using leave-one-out cross-validation: iteratively for each observation i, fit the model using only the observations $X_i - h \leq \mathbf{X} < X_i < c$ when $X_i < c$, and only the observations $c < X_i < \mathbf{X} \leq X_i + h$ when $X_i \geq c$.

Strengths & weaknesses

+ RDDs are similar to a local randomized experiment, and thereby require weak assumptions.
+ RDDs are all about finding "jumps" in the probability of treatment as we move along some X. They have much potential in economic applications, as geographic boundaries or administrative or organizational rules (e.g., program eligibility thresholds) often create usable discontinuities.
– They risk being underpowered.
– The parameter estimates are very "local", it may be hard to generalize from such a local result.

Fuzzy RD (imperfect compliance)

DGP At $X_i \geq c$ there is a jump, not in treatment assignment (D_i going from 0 to 1), but in the *probability* of treatment assignment $P[D_i{=}1 \mid X_i]$. The discontinuity $Z_i \equiv \mathbb{1}\{X_i \geq c\}$ becomes an instrumental variable for treatment status D_i. The DGP is represented in the causal graph below, where the brown arrows disappear as $X \to c$:



Estimand $$\beta_{\text{RD}} \equiv \frac{\lim\limits_{x \to c^+} \mathbb{E}[Y_i \mid X_i{=}x] - \lim\limits_{x \to c^-} \mathbb{E}[Y_i \mid X_i{=}x]}{\lim\limits_{x \to c^+} \mathbb{E}[D_i \mid X_i{=}x] - \lim\limits_{x \to c^-} \mathbb{E}[D_i \mid X_i{=}x]} = \ldots = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = c]}_{\textbf{LATE at the cutoff}}$$

Estimator Fuzzy RD leads naturally to a simple 2SLS estimation strategy. The 2SLS estimator $\hat{\beta}_{\text{2SLS}}$ is obtained through the two-step procedure:

$$\begin{aligned} \text{1st stage:} \quad & D_i = \delta + \gamma . Z_i + f(X_i - c) + u_i \implies \widehat{D}_i = \widehat{\mathbb{E}}[D_i | X_i] \\ \text{2nd stage:} \quad & Y_i = \alpha + \beta . \widehat{D}_i + f(X_i - c) + e_i \end{aligned}$$

As before, one can allow for treatment effects that change as a function of X_i by adding treatment-covariate interactions.

3.3 DiD, DiDiD, Event study

Repeated observations allow for controlling for unobserved confounders Repeated observations allow for controlling for all the characteristics — observed or not — that are constant over the given categorical dimension.
Say we have repeated observations for each individual i. Some characteristics stay constant for each individual (e.g., birth place, gender...). If before running our model, we subtract out of each observation the mean for that individual (for dependent and independent variables), we remove any variation explained by these characteristics. We say we "control for individual" as we get rid of all the variation explained by the individual, that is the *variation between* individuals. What's left is the *variation within* individuals.[a] We are now comparing that individual to themselves (over the dimension over which we have the repeated observations — typically time).[b]

There are two standard estimation methods to do this:

- Option #1: De-meaning manually: $Y_{it} - \bar{Y}_i = \alpha_0 + \beta(X_{it} - \bar{X}_i) + u_{it}$
  This quickly gets complicated if there are multiple dimensions along which to remove variation...
- Option #2: Adding "individual fixed effects", i.e., a set of binary indicators as explanatory variables, one for each individual: $Y_{it} = \alpha_i + \beta X_{it} + v_{it}$
  This is very easy to run with most statistical softwares. But it is estimating an intercept for each individual (even though we won't interpret them), which can be computationally intensive.

By including multiple sets of fixed effects, we would be controlling for multiple sets of unobserved confounders. What comparisons would then be averaged into β? For example, consider repeated cross-sectional data of groups $g = 1, ..., G$ over time $t = 1, ..., T$:

- w. group FEs: β = average of TEs that compare within-group peers across periods;
- w. year FEs: β = average of TEs that compare individuals within a period across groups;
- w. both ("two-way fixed effects"): β = average of TEs identified from variation within group and variation within year.[c]

[a] As the OLS estimator is a weighted average of treatment effects, with weights proportional to the conditional variance of treatment, the treatment effect estimated will weigh a lot more the units with large within variation.

[b] We could instead decide to have fixed effects for a higher level, e.g., city. We would then be comparing individuals in each city only to other individuals in that city.

[c] Each comparison is relative to what is expected given that group, and given that year. Note that this is not the same as "given that group that year" (such comparisons would rely on isolating variation within group-year, which would be obtained with group-by-year fixed effects). Here each of the "relative to" stands alone.

We'll consider here the common setting of a treatment assigned at a certain time, and units observed before and after the assignment, i.e., repeated observations over time. To estimate the causal effect of the event, we need a model to estimate the counterfactual value (the unit's outcome if the event had not occurred). The sections below present two different models of the counterfactual:

- DiD, DiDiD: when some units never get treated. We can use these control units to remove trends in Y in the treated. We identify effects from *within* and *between* variation.
- Event studies: when all units get treated. Assuming that a group's past value is a plausible counterfactual value, we identify effects from *within* variation only.

Assuming identification assumptions hold, estimation methods are very simple in the textbook setting (a binary D that only switches on, and is assigned at the same time for everyone). However, in more complicated settings (D discrete or continuous, switching off, staggered...), naive extensions of these methods may estimate quantities that are not the ATET!

DiD

DGP Treatment assignment or exposure is a function of two dimensions: group (treatment/control) and most commonly time (pre/post exposure).[26] We define the associated binary variables $G_i \equiv \mathbb{1}\{i \in \text{treatment group}\}$ and $P_t \equiv \mathbb{1}\{t \in \text{post period}\} \equiv \mathbb{1}\{t \geq \tau\}$.

- In a before/after comparison within the treatment group, the difference in Y could result from other changes that occur during the time period...
- In a treatment/control group comparison within the post period, the difference in Y could result from permanent differences between the groups...
- We can remove both biases by comparing the *change over time in* $\bar{Y}$ in the treatment group to the *change over time in* $\bar{Y}$ in the control group.

Identifying assumptions

(A1) Same or "parallel" counterfactual trends across groups: in the absence of treatment, both groups would have experienced the same *changes (pre → post) in outcomes* Y_{it}: $\mathbb{E}\big[Y_{i1}^0 - Y_{i0} \mid G_i{=}1\big] = \mathbb{E}\big[Y_{i1}^0 - Y_{i0} \mid G_i{=}0\big]$.

(A2) The group compositions do not vary over time.

Estimand $\beta_{\text{DiD}} \equiv \Big(\mathbb{E}\big[Y_{i1} \mid G_i{=}1\big] - \mathbb{E}\big[Y_{i0} \mid G_i{=}1\big]\Big) - \Big(\mathbb{E}\big[Y_{i1} \mid G_i{=}0\big] - \mathbb{E}\big[Y_{i0} \mid G_i{=}0\big]\Big) = \ldots = \underbrace{\mathbb{E}\big[Y_{i1}^1 - Y_{i1}^0 \mid G_i{=}1\big]}_{\textbf{ATET in post period}}$

Estimator The OLS estimator $\hat{\beta}_{\text{OLS}}$ of the following saturated regression consistently estimates β_{DiD}:

$$Y_{it} = \alpha + \beta_G G_i + \beta_P P_t + \boldsymbol{\beta}\, G_i P_t + e_{it}$$

In this 2-groups 2-periods design, β_{DiD} is equal to the treatment coefficient in a two-way fixed effect (TWFE) regression with group and period fixed effects: $Y_{it} = \lambda_G + \lambda_P + \boldsymbol{\beta}\, G_i P_t + u_{it}$ or $Y_{it} = \lambda_G + \lambda_t + \boldsymbol{\beta}\, G_i P_t + v_{it}$.

Best practices

- Support the assumption of parallel counterfactual trends by showing that pre-treatment trends coïncide (if we have data for multiple pre-periods). Estimate the following "event-study" or "dynamic TWFE" regression model by OLS, and check that the β_t for $t < \tau{-}1$ equal 0:
$$y_{it} = \lambda_G + \lambda_t + \sum_{t \neq \tau-1} \beta_t G_i \mathbb{1}\{t\} + e_{it}$$
- The regression above also enables us to look at whether the TE *accumulates* over time: $\beta_{t,t\geq\tau} \uparrow$ in t.
- If the composition of the groups changes over time, interact covariates X with P_t.
- As in any observational study, adjust for all other relevant pre-treatment variables.

Strengths & weaknesses

+ Repeated observations get rid of unobserved time-invariant confounders, creating comparable groups.
+ Pre-trends aren't a problem (unlike in event-studies) as long as that of the two groups are *parallel*.
+ Identification only requires repeated observations, so repeated cross-sectional data suffice, as long as the sample composition does not vary over time. Panel data satisfy this condition by construction.

DiDiD

DGP The treatment varies along a 3rd dimension or "subgroup", e.g., gender, space... We define the binary variable $S_i \equiv \mathbb{1}\{i \in \text{treatment in dim } \#3\}$.

[26] In the archetypical DiD setting, the second dimension is time, but it need not be. Data could be grouped by cohort (i.e., year of birth) or other characteristics.

Identifying assumptions

(A1) Same counterfactual trends across ~~groups~~ subgroups: in the absence of treatment, the difference in subgroups would have experienced the same *changes (pre → post) in outcomes*:

$$\mathbb{E}\left[Y_{i1}^0 - Y_{i0} \mid G_1, S_1\right] - \mathbb{E}\left[Y_{i1}^0 - Y_{i0} \mid G_1, S_0\right] = \mathbb{E}\left[Y_{i1}^0 - Y_{i0} \mid G_0, S_1\right] - \mathbb{E}\left[Y_{i1}^0 - Y_{i0} \mid G_0, S_0\right]$$

(A2) The subgroup compositions do not vary over time.

Estimand

$$\begin{aligned}\beta_{\text{DiDiD}} &\equiv \left[\left(\bar{Y}_{G_1S_1P_1} - \bar{Y}_{G_1S_1P_0}\right) - \left(\bar{Y}_{G_0S_1P_1} - \bar{Y}_{G_0S_1P_0}\right)\right] - \left[\left(\bar{Y}_{G_1S_0P_1} - \bar{Y}_{G_1S_0P_0}\right) - \left(\bar{Y}_{G_0S_0P_1} - \bar{Y}_{G_0S_0P_0}\right)\right] \\ &= \ldots = \underbrace{\mathbb{E}\left[Y_{i1}^1 - Y_{i1}^0 \mid G_i{=}1, S_i{=}1\right]}_{\textbf{ATET in post period}}\end{aligned}$$

Estimator The OLS estimator $\hat{\beta}_{\text{OLS}}$ of the following regression consistently estimates β_{DiDiD}:

$$Y_{it} = \alpha + \beta_G G_i + \beta_S S_i + \beta_P P_t + \beta_{GS} G_i S_i + \beta_{GP} G_i P_t + \beta_{PS} P_t S_i + \boldsymbol{\beta}_{\text{DiDiD}}\ G_i S_i P_t + e_{it}$$

Best practices

- A triple difference makes for a very specific control group. Before doing an DiDiD, one should explain why a double difference isn't satisfactory (i.e., why the control group in double difference isn't good enough s.t. the DiD assumption does not hold), and even the first difference.
- As in any observational study, adjust for all other relevant pre-treatment variables.

Strengths & weaknesses

+ A triple difference can difference out more confounding elements, hence it is harder to find confounders.
– It requires more data and variation.

Event study

DGP We want to estimate the causal effect of *an event*, which occurs at time τ and affects *all units* in the population, on some outcome Y. Treatment assignment is a function of the period (pre/post τ).

Identifying assumptions

(A1) Exogeneity (random timing): the event is unpredictable, and not a result of the outcome Y. We can then reasonably use the group's past value to construct its counterfactual post-event value.

(A2) The sample composition does not vary over time.

Estimand $\beta_{\text{ES}} \equiv \mathbb{E}\left[Y_{it} \mid t{=}\tau\right] - \mathbb{E}\left[Y_{it} \mid t{=}\tau{-}1\right] = \mathbb{E}\left[Y_{i,\tau}^1\right] - \mathbb{E}\left[Y_{i,\tau-1}\right] = \mathbb{E}\left[Y_{i,\tau}^1\right] - \mathbb{E}\left[Y_{i,\tau}^0\right] = \underbrace{\mathbb{E}\left[Y_{i\tau}^1 - Y_{i\tau}^0\right]}_{\textbf{ATET}}$

Estimator The OLS estimator $\hat{\beta}_{\text{OLS}}$ of the following regression (on a set of binary variables before and after the event, i.e., time fixed effects — omitting the period before the event to normalize it as 0) consistently estimates β_{ES}:

$$Y_{it} = \sum_{t=-K}^{\tau-2} \left[\beta_t\ \mathbb{1}\{t\}\right] + \boldsymbol{\beta}\ \mathbb{1}\{\tau\} + \sum_{t=\tau+1}^{L} \left[\beta_t\ \mathbb{1}\{t\}\right] + e_{it}$$

Best practices

- Plot/report all β_ts, to check that they are not changing up to the event. A change would suggest the presence of pre-trends, which making it hard to interpret the event (unless there is a sharp trend discontinuity) as they suggest some endogeneity of D.
- As in any observational study, adjust for all other relevant pre-treatment variables.

Strengths & weaknesses

– It is difficult to rule out other things changing at the same time, i.e., unobserved confounders.

⚠ Two-Way Fixed Effect estimators can be unreliable with heterogeneous TEs

Let's assume that the parallel trends assumption holds. We saw that in the canonical 2-groups 2-periods setting, β_{DiD} is equal to the two-way fixed effect (TWFE) estimator β_{FE}; both methods produce an unbiased estimate of the ATET. However, in designs with more variety in exposure to treatment (with many groups and periods, staggered treatment, treatment switching off, non-binary treatments...), the TWFE estimator may be biased **if TEs are not constant across groups or over time** (e.g., a policy becoming more or less effective). I.e., even with all confounders accounted for, β_{FE} is not robust to heterogeneity of TEs across groups or periods.

1. Source of the problem

Consider the TWFE regression model: $Y_{it} = \alpha_{g[i]} + \gamma_t + \beta_{\text{FE}} D_{g[i]t} + e_{g[i]t}$, for unit i in group g at time t. $\hat{\beta}_{\text{FE}}$ is a specific weighted sum of the ATE in each treated (g,t) cell, with each weight $\mathbf{w_{gt}}$ proportional to and of the same sign as $N_1(D_{gt} - D_{g.} - D_{.t} + D_{..})$ and $\sum \mathbf{w_{gt}} = 1$, such that in general, $\mathbb{E}[\hat{\beta}_{\text{FE}}] \neq \beta_{\text{ATET}}$ (de Chaisemartin and D'Haultfœuille, 2020):[27]

$$\beta_{\text{ATET}} = \mathbb{E}\left[\sum_{(gt):D_{gt}=1} \frac{N_{gt}}{N_1} \text{ATE}_{gt}\right], \qquad \mathbb{E}[\hat{\beta}_{\text{FE}}] = \mathbb{E}\left[\sum_{(gt):D_{gt}=1} \frac{N_{gt}}{N_1} \mathbf{w_{gt}} \text{ ATE}_{gt}\right]$$

✓ In the textbook case (D is binary, only switches on, and is assigned at the same time for everyone), $D_{gt} - D_{g.} - D_{.t} + D_{..}$ is constant across (g,t) cells, therefore $\hat{\beta}_{\text{FE}}$ is unbiased for the ATET.

✗ In more complicated settings, the $\mathbf{w_{gt}}$ vary; then heterogeneity in ATE_{gt} leads to a biased estimate. Some $\mathbf{w_{gt}}$ may even be negative (i.e., β_{FE} may not even identify a convex combination of TEs).[28]

– When D is staggered, in static TWFE regressions, $\hat{\beta}_{\text{FE}}$ is a weighted average of all possible 2-group, 2-period DiD estimators in the data, where each weight is a function of the sample size and the subsample variance of treatment (Goodman-Bacon, 2021).[29] Some of these DIDs misuse an early-treated group as control for a late-treated group, which may induce negative weights if the TE varies over time.

– When D is staggered, in dynamic TWFE or "event-study" regressions, the coefficient on a given lead or lag can be contaminated by effects from other periods, and apparent pretrends can arise solely from TE heterogeneity (Sun and Abraham, 2021).

Wooldridge (2021) highlights that the cause of the problem is not the TWFE estimator per se, but its misuse: it is applied to a restrictive model (which does not allow for heterogeneity in the TE).

[27] Where N_{gt} is the number of observations in cell (g,t); N_1 is the total number of treated observations; a dot subscript means the variable's average is taken over the given dimension; and while the original demonstration considers a binary treatment, the results "apply to any ordered treatment" (de Chaisemartin and D'Haultfœuille, 2022), s.t. $\text{ATE}_{gt} = (Y_{gt}^{D_{gt}} - Y_{gt}^0)/D_{gt}$. Precisely, $w_{gt} = \tilde{e}_{gt}/(\sum_{(gt):D_{gt}=1} \tilde{e}_{gt} N_{gt}/N_1)$, where $\tilde{e}_{gt}$ is the residual in the regression of D_{gt} on group and period FEs.

[28] This can even lead to a negative coefficient $\hat{\beta}_{\text{FE}}$ while the true ATEs are positive for everyone. Ex: $1.5 \times 1 - 0.5 \times 4 = -0.5$.

[29] OLS will give more weight to subgroups where the FE-adjusted treatment dummy varies more. As a result, the timing of a unit's treatment will determine its weight in the regression. If i is treated very early or very late, then it will have very little variation in treatment across the period (is 0 almost the whole time, or 1 almost the whole time) and so will receive little weight. Note that weighting by treatment-variance is how OLS handles heterogeneity all the time — see section 2.1.2.

2. Alternatives

de Chaisemartin and D'Haultfœuille (2022) summarizes the fast-growing literature on this problem, and highlights:

- Diagnosis tools
  - The `twowayfeweights` command (in R and Stata) computes the weights $\frac{N_{gt}}{N_1}\mathbf{w_{gt}}$.
  - The `bacondecomp` command (in R and Stata) computes the DID estimators and weights entering in β_{FE}, in the case of a binary staggered D.
- Alternative estimators
  Ex: Wooldridge (2021) proposes an "extended TWFE" approach (based on the random-effects Mundlak estimator) which notably interacts the treatment indicator with time and/or group-time dummies to allow TEs to change across groups or periods.

More generally, since the source of the 'problem' is the heterogeneity in TE, we should be thinking about how to allow for heterogeneity in our model.

3.4 SCM

Summary of common identification methods

	Source of identification & identifying assumptions	**Estimand β & corresponding TE**	**Chosen estimator $\hat{\beta}$**	**Strengths / Weaknesses**
RCT	(A) independence	$\beta_{\text{RCT}} \equiv \mathbb{E}[Y_i \mid D_i{=}1] - \mathbb{E}[Y_i \mid D_i{=}0] = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0]}_{\textbf{ATE}}$	$\hat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \alpha + \boldsymbol{\beta}\, D_i + e_{it}$. **Consistent** and **unbiased**.	+ Random assignment structurally guarantees (A) $\Longrightarrow$ RCT = *"gold standard"*
IV	Id. from the exogenous variation in D induced by Z. (A1) independence (A2) exclusion restriction (A3) relevance (A4) monotonicity	$\beta_{\text{IV}} \equiv \frac{\text{cov}[Y_i, Z_i]}{\text{cov}[D_i, Z_i]} = \ldots$ $= \frac{\mathbb{E}[Y_i \mid Z_i{=}1] - \mathbb{E}[Y_i \mid Z_i{=}0]}{\mathbb{E}[D_i \mid Z_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0]}$: *"Wald estimand"* $= \ldots = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid D_i^1{=}1, D_i^0{=}0]}_{\textbf{LATE, compliers}}$	$\hat{\beta}_{\text{W}} \equiv \frac{\widehat{\text{cov}}[Y_i, Z_i]}{\widehat{\text{cov}}[D_i, Z_i]} = \ldots =$ numerically equivalent to $\hat{\beta}_{\text{2SLS}}$ **Consistent, biased**, but bias $\downarrow$ with strength of Z_i.	+ compelling identification strategy – strong assumptions – less efficient than $\hat{\beta}_{\text{OLS}}$ if instrument is weak
sharp RD	Id. from a discontinuous treatment assignment based on a cutoff in X. (A1) local continuity (A2) relevance	$\beta_{\text{RD}} \equiv \lim_{x \to c^+} \mathbb{E}[Y_i \mid X_i{=}x] - \lim_{x \to c^-} \mathbb{E}[Y_i \mid X_i{=}x]$ $= \ldots = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid X_i{=}c]}_{\textbf{LATE, at the cutoff}}$	$\hat{\beta}_{\text{OLS}}$ of the regression $Y_i = \alpha_l + \boldsymbol{\beta} D_i + f(X_i{-}c) + e_i$, with choice of $f()$: – local linear regression – polynomial regression **Consistent, biased**, bias $\uparrow$ w. bandwidth	+ akin to a local randomized experiment + weak & testable assumption – risks being underpowered – low external validity
fuzzy RD	Id. from a discontinuous *P(treatment assignment)* based on a cutoff in X. (A1) local continuity (A2) relevance	$\beta_{\text{RD}} \equiv \frac{\lim_{x \to c^+} \mathbb{E}[Y_i \mid X_i{=}x] - \lim_{x \to c^-} \mathbb{E}[Y_i \mid X_i{=}x]}{\lim_{x \to c^+} \mathbb{E}[D_i \mid X_i{=}x] - \lim_{x \to c^-} \mathbb{E}[D_i \mid X_i{=}x]}$ $= \ldots = \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = c]}_{\textbf{LATE at the cutoff}}$	$\hat{\beta}_{\text{2SLS}}$	
DiD	(A1) same counterfactual trends across groups (A2) same group compositions over time	$\beta_{\text{DiD}} \equiv \left(\bar{Y}_{G_1P_1} - \bar{Y}_{G_1P_0}\right) - \left(\bar{Y}_{G_0P_1} - \bar{Y}_{G_0P_0}\right)$ $= \ldots = \underbrace{\mathbb{E}[Y_{i1}^1 - Y_{i1}^0 \mid G_i{=}1]}_{\textbf{ATET}}$	$\hat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \boldsymbol{\beta}\, G_i P_t + \lambda_G + \lambda_P + e_{it}$ **Consistent**.	+ rules out unobserved time-invariant confounders
DiDiD	(A1) same counterfactual trends across subgroups (A2) same subgroup compositions over time	$\beta_{\text{DiDiD}} \equiv \Big[\left(\bar{Y}_{G_1S_1P_1} - \bar{Y}_{G_1S_1P_0}\right) - \left(\bar{Y}_{G_1S_0P_1} - \bar{Y}_{G_1S_0P_0}\right)\Big]$ $- \Big[\left(\bar{Y}_{G_0S_1P_1} - \bar{Y}_{G_0S_1P_0}\right) - \left(\bar{Y}_{G_0S_0P_1} - \bar{Y}_{G_0S_0P_0}\right)\Big]$ $= \ldots = \underbrace{\mathbb{E}\left[Y_{i1}^1 - Y_{i1}^0 \mid G_i{=}1, S_i{=}1\right]}_{\textbf{ATET}}$	$\hat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \boldsymbol{\beta}\, G_i S_i P_t + \lambda_{GS} + \lambda_{GP} + \lambda_{PS} + e_{it}$ **Consistent**.	+ differences out more confounding elements than in DiD, so harder to find a confounder – requires more variation
Event-study	(A1) random timing of event (A2) same sample composition over time	$\beta_{\text{ES}} \equiv \mathbb{E}\left[Y_{it} \mid t{=}\tau\right] - \mathbb{E}\left[Y_{it} \mid t{=}\tau{-}1\right]$ $= \ldots = \underbrace{\mathbb{E}\left[Y_{i\tau}^1 - Y_{i\tau}^0\right]}_{\textbf{ATET}}$	$\hat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \boldsymbol{\beta}\, \mathbb{1}\{\tau\} + \sum_{t \neq \{\tau-1, \tau\}} [\beta_t \mathbb{1}\{t\}] + e_{it}$ **Consistent**.	+ flexible – difficult to rule out unobserved confounders
SCM				

4 Analysis stage: Steps for stronger causal inferences

4.1 Identification strategies only provide so much

Recall the core motivation for identification strategies:

> We look for identification strategies that suggest that an independence assumption holds, as:
>
> - if IA, the regression of Y on D gives an unbiased estimate of the ATET *(e.g., in an RCT)*;
> - if ~~IA~~ CIA + we know the correct functional form $f()$ w.r.t. X, the regression of Y on D and $f(X)$ gives an unbiased estimate of the ATET;
> - in either case, if we instrument D by a valid Z, IV regression gives an unbiased estimate of a LATE.

All that identification strategies buy us is the above. This is actually very limited, in at least 3 major ways:

1. In observational studies, we always have at best a ~~IA~~ CIA. Then unbiased estimation of the ATET relies on correctly specifying the functional form w.r.t. X. But we don't ever know this $f()$ for sure.[30] So we don't want to have to rely on $f()$. Then we must strive to avoid areas of imperfect overlap in our data (there, we are forced to rely on model specification instead of direct support from the data, so inferences are vulnerable to model misspecification...). I.e., assuming the CIA holds, accurate estimation is still not guaranteed, but comes down to proper modeling and the extent to which the model is forced to extrapolate beyond the support of the data.

   This is related to the literature on "doubly-robust" estimators (Hill, 2011, section 2).

2. An unbiased estimator $\hat{\theta}$ just means that its distribution $f_{\hat{\theta}}$ (over possible trials for the given sample size) is correctly centered (around the true value of the estimand θ); it does not guarantee that its realization for any particular study will be close to that center value — especially with a small sample size. We might therefore want to:

   (a) adjust as much as possible for potential imbalance between the groups, using pre-treatment data;
   (b) consider another property: efficiency (i.e., reduce the width of the estimator's distribution).

3. We obtain an estimate of the ATET, but what knowledge are we generating from that? Reduced forms are generally — this document is no exception — motivated by having set the RCT as gold standard. In an RCT, the treatment variable represents an intervention, so the average effect of that intervention might very well be the knowledge desired. However, in other contexts, estimating the magnitude of an effect without identifying its underlying mechanisms[31] might be considerably less informative (e.g.g, the impact of climate extremes on social instability).

This section suggests what can be done at the analysis stage (i.e., post-design, given a fixed dataset) to try to counteract these limitations, and generate more insightful inferences. Specifically:

1. pre-estimation: restructuring the data to improve overlap;
2. in estimation: adjusting as much as possible for potential imbalance, and allowing for TE heterogeneity;
3. post-estimation: checking assumptions and considering external validity.

4.2 *Pre-estimation:* Restructuring

Causal inference requires the units in the treatment group to be comparable to those in the control group w.r.t. confounders X. There are two forms of departures from comparability:

[30] One way to avoid possible misspecification would be to saturate the model, i.e., discretize each variable in X using indicator variables, and include a separate parameter for every possible combination of values of this set of regressors. This is rarely tractable in practice, notably with continuous X.

[31] As discussed in the following subsection, adjusting for "intermediate outcomes" to estimate so-called mediating effects will bias the treatment effect estimate.

- Incomplete overlap: the *support* of the distribution of X differs across the groups. Some observations have no empirical counterfactuals.
  ↪ In these zones of no overlap, the model is forced to extrapolate, and inferences are based entirely on modeling assumptions instead of data.
- Imbalance: the *shape* of the distribution of X differs across the groups (e.g., different means, same mean but different skews).
  ↪ The simple difference of group averages is not, in general, a reliable estimate of the ATET.

Restructuring to balance the observed confounders The less the treatment and control groups have overlap and balance w.r.t. confounders X, the more our inferences rely on the model instead of on data, and so aren't robust to model misspecification. On the contrary, if the distributions of X are similar across the groups, then, even if we misspecify the form of the relationship, we should still get a reasonable estimate of the TE (Gelman et al., 2020). To alleviate this concern of needing to specify the model correctly, we can *restructure* our sample prior to analysis, namely match groups to exhibit balance and overlap w.r.t. the confounders X (i.e., make the sample resemble one from a randomized trial: $D \perp\!\!\!\perp X$). As the estimand of interest is the ATET, we want our analysis sample to be representative of the *treatment* group, so we will keep our treatment group intact and restructure the control group to look like the treatment group.

⚠ *Matching provides more overlap and balance, not identification. For matching to be able to capture by itself a causal effect, all the difference between the groups would need to be captured by observed X. This assumption of "selection on observables" is very strong, and not testable. Therefore matching is not an alternative to a design-based method.*[32] *We need exogenous variation to believe the CIA. Matching is an adjustment strategy, not an identification strategy.*

With[33] $\begin{cases} \textit{(i) CIA} \\ \textit{(ii) balance \& overlap w.r.t. } X \end{cases} \implies$ *the difference in $\bar{Y}$ is an unbiased estimate of the ATET.*

As an adjustment strategy, one can nonetheless use matching in two different ways:

- **In place of regression: matching as estimation method**
  The regression with controls estimand or the covariate-matching estimand are two different ways to balance the Xs (Angrist and Pischke, 2008). In practice, the matching estimator is computed by making comparisons for cells with the same X values, computing the difference in their Ys, and averaging these differences in some way.
  However, while neither the regression nor the matching *estimands* give any weight to covariate cells that don't have both treated and control observations (i.e., both estimands impose common support), the regression and matching *estimators* use modeling assumptions that implicitly involve extrapolation across cells, so cells without both treated and control observations can end up contributing to the estimates by extrapolation. Using matching as an estimation method therefore does not resolve the concerns with lack of overlap. Estimating the standard errors of matching estimates is also not straightforward.
- **On top of regression: matching as preprocessing method**
  2-step process: do matching to get comparable groups, and then do regression for further adjustment and for modeling interactions. Matching as a nonparametric preprocessing procedure is used to restructure the original sample before statistical analysis, to reduce reliance on the parametric assumptions of the subsequent regression model (Gelman et al., 2020; Ho et al., 2007).

Common distance metrics One can match units rather easily with one continuous confounder X (choose for each treated unit the control unit with the closest value of X), or even one binary X_1 and one continuous X_2 (e.g., stratify within subgroups defined by X_1 and then match on X_2 within each subgroup). But it

[32] Methods in which a feature in the setting approximates a randomized experiment, and we fit a model that adjusts for potential confounders: RDs, IVs... (the methods described in the previous section).

[33] In other words, identification strategies and econometrics (matching, regression) are complements in the production of causal estimates. Some independence assumption is needed to give estimates (whether matching estimates or regression coefficients) a causal interpretation.

quickly gets complicated with more confounding covariates. One alternative is to define a univariate distance metric between observations as a function of the Xs, and match each treated unit to its nearest control unit:

- **Mahalanobis distance**
  We define a distance metric that can include multiple dimensions of "closeness" between observations: $d_{ij} \equiv \sqrt{(X_i - X_j)'\Sigma_X^{-1}(X_i - X_j)}$, where Σ_X is the sample covariance matrix. This distance metric is scale-invariant and accounts for the correlation structure of the Xs.

- **Propensity score**
  We can reduce the dimensionality to 1 by computing a unit's predicted probability of getting treated or "propensity score" $\hat{p}_i$ from the Xs, and use as distance metric $d_{ij} \equiv |\hat{p}_i - \hat{p}_j|$.

  The appeal of the propensity score $p(X)$ is that if the Xs included in the propensity score model are sufficient to satisfy ignorability, then $p(X)$ is also sufficient to satisfy ignorability. I.e., appropriate conditioning on $p(X)$ (for instance by matching or weighting on functions of it) is sufficient to estimate an unbiased TE. If $Y^1, Y^0 \perp\!\!\!\perp X$, then $Y^1, Y^0 \perp\!\!\!\perp p(X)$.

  Algorithm for propensity score matching:

  1. Propensity score model: fit a logistic regression of D_i on $\{X_i\text{s}\}$, and predict $\hat{p}_i \equiv \text{P}(D_i = 1|X_i)$
  2. Match each treated unit to its nearest control unit(s) using $\hat{p}_i$. Choose matching algorithm: with/without replacement;[a] coarse (stratify the sample into quintile blocks of $\hat{p}_i$)...
  3. Diagnose: assess balance & overlap. If balance is inadequate, redo steps 1-3, trying a less parsimonious model (add interactions, higher order terms of covariates...), a less coarse matching algorithm...
     - Balance: compare the distribution of each X across the groups, before vs after matching.
     - Overlap: plot overlapping histograms w.r.t. the estimated propensity score.
  4. Estimate the ATET using the restructured data. As aforementioned, one can elect to either:
     - estimate the ATET as a weighted difference in means (i.e., compute a matching estimator). Example: Almond et al. (2005) presents both an OLS and a matching estimator.
     - fit a regression model
       * on D, $\hat{p}$ and $D \times \hat{p}$. See Doug's "Algorithm for Estimating the Propensity Score" pdf — from Ken Chay's 2001 UC Berkeley econometrics class
       * on D and all Xs using the restructured data. Gelman et al. (2020, ch. 20): *"This gives us an additional chance to adjust for differences in distributions of X that typically remain between the groups (to decrease bias and increase efficiency). The overlap and balance created by the matching should make this model more robust to potential model misspecification; that is, even if this model isn't quite right (for example, excluding a key interaction,or assuming linearity when the underlying relation is strongly nonlinear) our coefficient estimate should still be close to correct, conditional on ignorability being satisfied."*[b]

  [a] Matching w. vs without replacement is akin to a bias-variance trade-off: matching with replacement should yield better matches on average, therefore better balance and less biased TE estimates; however, it can result in over-using certain units or ignoring other close matches, i.e., missing out on important information in the data, and potentially increase the variance of the estimates.

  [b] ⚠ The standard errors from this regression are not technically correct, as: (1) matching induces correlation among the matched observations — the regression model, however, if correctly specified, should account for this by including the variables used to match; (2) the fact that the propensity score has been estimated from the data is not reflected in the calculations (Gelman et al., 2020, p. 404).

4.3 *Estimation:* Regression controls and interactions

Good/bad regression controls We saw that to recover an unbiased TE estimate, we must adjust for all *confounding variables* (variables that correlated with both D and y); adding confounders as covariates is part of the identification strategy. Separately from the identification strategy, which other covariates should we include, i.e., adjust for?

- 👍 Adjusting for *pre*-treatment covariates that have a strong association with y can increase the efficiency, i.e., precision, of the estimate. They will reduce the residual variance (the unexplained variation in y) and thereby lower the standard error of the regression estimates, even though they are uncorrelated with D. This applies also to data from a completely randomized experiment.
- 👎 Do not adjust for *post*-treatment variables! Covariates that may be affected by the treatment or that are highly correlated with it may introduce bias.

We must however be careful to avoid overfitting, and when having a large number of covariates, of finding a way to choose among them. Ways/criteria to penalize complexity in linear regressions and variable selection:

- Familiar: adjusted R2
- Elastic net regression: minimizes the sum of squared residuals plus a penalty term, to choose the regression coefficients $\{\beta_p\}$:

$$\{\beta_p\} = argmin SSR + \lambda \sum_p \left[(1-\alpha)|\beta_p| + \alpha|\beta_p|^2 \right]$$

  It overcomes the limitations of LASSO. If $\lambda = 0$, this is OLS; if $\alpha = 0$, this is LASSO.
  Suresh Naidu suggests reporting robustness of estimated treatment effect of interest to different values of λ with LASSO; rather than arbitrary author-curated specifications across various columns of a table.

TE heterogeneity We expect some heterogeneity in the treatment effect, we might therefore want to look into it. In particular, if we have adjusted for a pre-treatment covariate X that has a large estimated effect, it is natural to look at how the TE varies with the level of that X (Gelman et al., 2020). A simple way of doing this is by interacting the treatment D with X.[34]

4.4 *Post-estimation:* Supporting assumptions & Predictions

4.4.1 Diagnosis tests of modeling assumptions

4.4.2 Falsification tests of identifying assumptions

One can never directly *test* the identifying assumptions, i.e., prove that they hold. But one can do falsification analyzes that will either increase or decrease our confidence in them — and thus support the internal validity of the study. These are often referred to as "falsification" or "placebo" tests.

Balance table Causal inference rests upon the assumption that the treatment and control groups are comparable to some extent — eventually, conditional on some covariates. In the 'gold standard' RCT, the identifying assumption is random assignment. If treatment was indeed randomly assigned, then the sample means of explanatory variables should be the same across the treatment and control groups *in expectation*. RCT papers therefore typically show a "balance table" of sample means of the Xs by group.[35] Doug Almond's advice: even in observational settings, *always* show a balance table. Andrew Gelman's advice: to show balance in averages of X across groups, plot the standardized and absolute differences in mean values for the continuous and binary Xs, respectively.

[34] Consider centering X, s.t. the treatment coefficient represents the TE for individuals with the mean X score for the sample.

[35] ⚠ In any study where treatment was actually randomly assigned, such as in an RCT, one should show for each observable X the difference of sample means between the two groups, but not a t-test of whether it is significantly different from zero. A t-test tells us how significant the difference of group means is, i.e., whether that difference could have happened by chance. The randomization already guarantees unbiasedness, i.e., that any difference observed would have happened by chance. t-tests in this context are therefore conceptually unsound. Hayes and Moulton (2017) explain that *"the point of displaying between-arm comparisons is not to carry out a significance test, but to describe in quantitative terms how large any differences were, so that the investigator and reader can consider how much effect this may have had on the trial findings."* One can document (im)balance through some distance measure, for example the normalized difference $\Delta X \equiv (\bar{X}_1 - \bar{X}_0)/\sqrt{S_0^2 + S_1^2}$ (where S_0^2 and S_1^2 are the sample variances of X in the control and treatment groups, respectively) (Imbens and Wooldridge, 2009, eq. 3).

Other tests In observational studies, the general approach to support core identifying assumptions is to demonstrate that the specification does not find an effect when it indeed "should not" exist, e.g., by looking at an outcome which should not be affected under the identifying assumption. If the analysis picks up an effect where there isn't one, it suggests that the identifying assumption is violated, a confounder is probably driving the relationship.

- **IV** The two main identifying assumptions can be tested:
  - relevance (Z is strongly related to sorting into treatment D): directly observable in the 1st stage;
  - exclusion restriction (Z isn't correlated with Y through some pathway other than D). The ideal falsification test is to estimate the reduced form effect of Z on Y in a situation where Z can't affect D. Finding an effect means Z affects Y through another channel than D, falsifying the exclusion restriction. One can use an alternative population or an alternative outcome, that can't be affected by the treatment but would be by potential confounders (unobserved characteristics correlated with Z and Y).
- **RD** The two main identifying assumptions can be tested: `https://mixtape.scunning.com/regression-discontinuity.html?panelset2=r-code3#mccrarys-density-test`
  - continuity or "local randomization" (all other factors determining Y evolve "smoothly" w.r.t. Z). Test: do other covariates jump at the cutoff c? Estimate the same model, but using covariates instead of Y, and plot the observations and the fitted curves. If none do, we can assume the unobservables don't either.
  - relevance (discontinuity in the dependence of D on Z: $D = \mathbb{1}\{Z \geq c\}$). Test: do jumps occur at placebo cutoffs $\tilde{c}$?
- **DiD** The two main identifying assumptions can be tested:
  - same counterfactual trends across groups. Tests: compare trends in the pre-period; use an alternative outcome that shouldn't be affected by the treatment; use an alternative control group (the estimated effect should be the same); move the event to points earlier in time (falsely assume that the onset of treatment occurs before it actually does), if the estimated treatment effect is no longer be statistically significant (i.e., is statistically indistinguishable from 0.), suggests that the observed change is more likely due to the treatment (event) than to some alternative force.
  - same group composition over time. Panel data satisfies this assumption by definition, but if we have instead repeated cross-sectional data, we can estimate covariate balance regressions.

Examples

DiD Linden and Rockoff (2008): *What is the hedonic price function for the local disamenity of crime risk (i.e., individuals' valuation of crime risk)?* Y = property value, D = a registered sex offender moves in nearby.
Falsification test of the "same counterfactual trends" assumption (if the prices of houses in offender areas are trending over time differently than the other houses in their neighborhood, they would estimate a spurious negative "impact" of the offender's arrival): the authors estimate the DiD model using false arrival dates (2-3 years prior to an offender's actual arrival), and find no effect.

Note: Falsification tests are different from robustness checks, which consist in estimating alternative specifications that test the same hypothesis.

4.4.3 Mechanisms & External validity

Validity of a statistical analysis

- **Internal validity** = the extent to which the causal effect *in the population being studied* is *properly identified*. It is determined by how well the study can rule out alternative explanations for its findings.
- **External validity** = the extent to which the inferences can be generalized to other populations and settings.
  ⚠ Even in randomized trials, the experimental sample often differs from the population of interest. If participation decisions are explained by observed variables, such differences can be overcome by reweighting, but participation may depend on unobserved variables.

5 Presentation

5.1 Characterizing the empirical strategy

The empirical strategy for any econometric analysis aiming for causal inference should contain — to some degree, explicitly — the following items:

1. Research question — *What causal effect of interest are we trying to estimate?*
2. Ideal experiment — *What ideal experiment would capture the causal effect?*
3. Identification strategy — *How are the observational data at hand used to make comparisons that approximate such an experiment? Specifying notably: the identifying assumptions, what makes them satisfied, the specific effect estimated (ATET, LATE...).*
4. Estimation method (incl. assumptions made when constructing standard errors).
5. Falsification tests that bring confidence in the identifying assumptions.

All these items can be characterized before opening the dataset.

5.2 Putting the paper in perspective

In addition to the paper's empirical strategy, one may want to discuss:

- Contributions to the literature on the topic or research question
- Methodological contributions
- Internal validity of the statistical analysis
  Are the identifying assumptions plausible (are there stories under which the assumptions would not hold?) Could there be measurement error? Are there unexplained results?
- External validity of the statistical analysis
  - w.r.t. policy: is there a gap between policy questions and the analyses performed?
  - w.r.t. the literature: how does the paper account for its results compared to other results in the literature?
  - w.r.t. other settings: are the results generalizable to other populations and settings?

6 Other branches of causal modeling

6.1 *Which uncertainty matters?* Randomization inference (RI)

"In randomization-based inference, uncertainty arises naturally from the random assignment of the treatments, rather than from the hypothesized sampling from a large population." (Athey and Imbens, 2017)

The inference techniques we commonly use in regression analysis correspond to *sampling-based* inference. They consider variation in sampling: the uncertainty about population parameters is induced by random sampling from the population. These methods ask: *What would have occurred under a different random sample than the one sampled?*

In causal inference studies, there is also another type of variation at play: variation in *assignment of treatment*, i.e., *design-based* uncertainty corresponding to what the regression outcome would have been under alternative randomizations of treatment assignment. In "Randomization Inference", introduced by Fisher (1935), the basis for inference is the distribution induced by the randomization of the treatment allocation. One takes *"a design-based perspective where the properties of the estimators arises from the stochastic nature of the treatment assignment, rather than a sampling-based or model-based perspective where these properties arise from the random*[36] *sampling of units from a large population in combination with assumptions on this population distribution"* (Athey and Imbens, 2022). One asks: *What would have occurred under a different random assignment of treatment among units than the assignment observed?*

Application to hypothesis testing Both sampling-based and design-based inference follow the same approach to hypothesis testing: we formulate a null hypothesis that represents a fact about the data we'll try to refute. In causal inference, it is generally a hypothesis of no effect. We then derive a test statistic T s.t. when H_0 is true, T has a specific distribution, and we look at where the value of T *for our observed data* $\hat{T}_{\text{obs}}$ lies within that distribution. The furthest in the tails, the less likely these observed data were under the null hypothesis, therefore the higher the confidence against it.

In randomization inference, considering the *sharp* null hypothesis of no effect *for any unit*,[37] we can simply use β as the test-statistic and obtain its empirical distribution under H_0. Indeed:

- If there is no effect for any unit, then a unit's potential outcomes are identical: the observed outcome is also the counterfactual. Under H_0, our data therefore represent the outcomes of all possible experiments.
- If we construct all possible random assignments, estimate $\hat{\beta}$ for each, the resulting distribution of $\hat{\beta}$ is therefore *the* reference distribution under H_0.
- We look at where our actual $\hat{\beta}_{\text{obs}}$ falls in the reference distribution; if in the tails, e.g., such that only 2% of all random assignments produce a $\hat{\beta} \geq \hat{\beta}_{\text{obs}}$, our one-tailed p-value is 0.02.

In practice: simulation When *all* possible random assignments can be simulated, the reference distribution is known, thus RI produces *exact* p-values. In practice, the number of possible assignments is generally huge, so we don't simulate all of them but many, to approximate the reference distribution, and compute approximate p-values. We repeat a large number of times (e.g., 10000) the following procedure:[38]

1. Re-assign treatment randomly, i.e., draw from the "randomization set"[39] (respecting the structure of the original assignment mechanism, e.g., within strata), thus generating fake treatment statuses.

[36] At this point the term 'randomization' might seem confusing, as *both* approaches assume and build inference from randomness: in the traditional approach, that of the *sample*; in the design approach, of the *treatment assignment*. There is a subtle difference: in the first the sample isn't *randomized* but simply *random*, i.e., taken randomly, whereas in the second, because assignment is made in a random fashion, the resulting treatment is first randomized, and therefore random. RI *is* aptly named.

[37] Note that this is substantially different from the usual null hypothesis in sampling-based inference of *no average effect*.

[38] RI is a simulation approach, like Bootstrap, however Bootstrap considers variation from sampling. A Bootstrap procedure resamples observations from our actual sample (which is fair, as we assumed it was representative of the population), with replacement, to simulate how *sampling* variation would affect our results.

[39] (Rubin, 1974) defines the "randomization set" as "the set of allocations that were equally likely to be observed given the randomization plan". Ex: for a completely randomized experiment of $2N$ trials, where N is assigned to each treatment arm, there are $\binom{2N}{N}$ possible allocations.

2. Estimate the regression model using these fake treatments, and store the $\hat{\beta}$s.

We obtain a distribution for the $\hat{\beta}$s.

Sampling-based inference	Randomization inference
H_0, H_a	
H_0: No average effect: $\mathbb{E}[Y_i^1] - \mathbb{E}[Y_i^0] = 0$ H_a: An average effect: $\mathbb{E}[Y_i^1] - \mathbb{E}[Y_i^0] \neq 0$	H_0: "Sharp" no effect: $Y_i^1 - Y_i^0 = 0, \forall i$ H_a: $\exists i$ s.t. $Y_i^1 - Y_i^0 \neq 0$
T & distribution of T under H_0	
$T \equiv (\beta - 0)/\text{SD}(\beta)$, $\hat{T} = \hat{\beta}/\text{SE}(\hat{\beta})$ Under H_0, the distribution of T across all random samples converges (as $n \to \infty$) to a known distribution: Student's `t`. $\longrightarrow$ We compute the parameters of this distribution. $\longrightarrow$ The *asymptotic* distribution of $\hat{T}$ (across all random samples) = the "sampling distribution under H_0".	$T \equiv \beta$, $\hat{T} = \hat{\beta}$ Under H_0, how the treatment was randomly assigned wouldn't change the observed outcomes; but it would change the value of $\hat{T}$. $\longrightarrow$ We compute $\hat{T}$ for all possible random assignments. $\longrightarrow$ The *exact* distribution of $\hat{T}$ (across all random assignments) = the "reference distribution under H_0".
2-sided p-value $= \mathbf{Pr}\left[\text{observing a } \hat{T} > \lvert\hat{T}_{\text{obs}}\rvert\right]$ **under** H_0 $=$ **share of the distribution that is** $> \lvert\hat{T}_{\text{obs}}\rvert$	
$= \Pr\left[\textit{the observed difference between groups would have been observed}\right]$ if they had been drawn from underlying sampling frames with no mean difference.	$= \Pr\left[\textit{the observed difference between groups would have been observed}\right]$ if the TE were in fact 0 for every subject.
$\Longrightarrow$ Given e.g. a rejection threshold $\alpha = 0.05$, the test will erroneously reject H_0 < 5% of the time	

Why choose randomization-based inference instead of sampling-based inference?

- Conceptually, there is sometimes no true sampling variation to speak of. Suppose we observed the universe of y outcomes, then there is no sampling from a large population, making sampling-based p-values meaningless, SE = 0.[40] Regardless, the core uncertainty within a causal study is not solely driven by the universe of possible samples, but also by the universe of possible treatment assignments.
- RI is not confined to large samples. As we don't have to appeal to the asymptotic properties of an estimator, it allows us to make inferences about causal effects even in settings where assuming an infinite number of treatment units may not be credible.
- RI is not confined to normally distributed outcomes. The method can be applied to all sorts of outcomes, such as counts, durations, ranks (Gerber and Green, 2012, p.63).
- RI salvages inference with particular clustered designs
  - *Small number of assignment clusters:* When the number of clusters is small, cluster-robust standard errors are downwardly biased. RI circumvents this problem as the reference distribution is calculated based on the set of possible clustered assignments, which takes into account the sampling variability associated with clustered assignment.
  - *Assignment clusters without well-defined boundaries:* if the assignment clustering isn't within well-defined boundaries, one can't rely on common methods to estimate correct standard errors (clusters can't be defined; other sandwich-type covariance matrix estimators require additional modeling assumptions...). Ex: weather variables such as rainfall are often used as a strategy for causal inference, as rainfall shocks are as-if randomly assigned. However, the assignment of rainfall is highly correlated across space in an unformalizable structure. Cooperman (2017) uses national draws of historical rainfall patterns as potential randomizations, allowing her to preserve patterns of spatial dependence while remaining agnostic about the specific form of the clustering.[41]

[40] While it is indeed possible to observe the value of a variable for all the units in a population (e.g., the eye colors of the 50 U.S. senators), one rarely observes all the possible range of values that units could have taken. Thinking of that universe of values as the relevant population alleviates the conceptual concern.

[41] Note that the use of historical data is disputable if climate change changes the distribution across years.

- Apparently RI is somewhat more robust to the presence of leverage in a few observations.
  Young (2019) collected over fifty experimental (lab and field) articles from the American Economic Review, American Economic Journal: Applied, and American Economic Journal: Economic Policy. He then reanalyzed these papers, using the authors' models, by dropping one observation or cluster and reestimating the entire model, repeatedly. He found that with the removal of just one observation, 35% of 0.01-significant reported results in the average paper can be rendered insignificant at that level, 16% of 0.01-insignificant reported results can be found to be significant at that level. In the typical paper, randomization inference found individual treatment effects that were 13 to 22 percent fewer significant results than what the authors' own analysis had discovered.
  Young, Alwyn. 2019. "Chanelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." Quarterly Journal of Economics 134 (2): 557–98.

Limitations

- RI is a method for hypothesis testing — not for constructing confidence intervals!
  ⚠ The "reference distribution under H_0" does not give confidence intervals for $\hat{\beta}$. It is instead the set of all possible estimated values of $\hat{\beta}$ when the true $\beta = 0$. This does not represent our statistical uncertainty about $\hat{\beta}$, it only enables us to compute p-values for the sharp null hypothesis of no effect.
  - ↪ Is RI even worth it then? After all, the 2-way binary approach to statistical hypothesis testing, based on the NHST falsificationist paradigm and the formulation of binary statements of 'statistical significance' from a p-value threshold, is heavily criticized...
- RI may also be used for construction of confidence intervals, but this application requires additional assumptions.
  - Rosenbaum (2002, p.45) proposes a method by "inverting" the hypothesis test.
  - Gerber and Green (2012, p.67) proposes a simpler — but less accurate — method, and argues that the two methods tend to produce similar results, especially in large samples.
  - Barrios et al. (2012, eq. (4.2)) gives the exact conditional (randomization-based) variance of $\hat{\beta}$ (in the univariate linear regression model of Y on D) under the assumption of a homogeneous treatment effect, based on Neyman (1923) (unfortunately, the proof is omitted):

$$\mathbb{V}[\hat{\beta}_{\text{OLS}}|e] = \frac{N}{N_0 N_1 (N-2)} \sum_i (e_i - \bar{e})^2, \quad \text{where } N_1 \equiv \sum_i D_i,\ N_0 \equiv N - N_1$$

6.2 Structural Equation Models (SEMs)

Structural Equation Models are probabilistic models that unite multiple predictor and response variables in a single causal network.

SEMs are increasingly popular in ecological research. They are often represented using path diagrams, a.k.a. directed acyclic graphs (DAG), where arrows indicate directional relationships between observed variables.

Implicit assumptions — what separate SEMs from traditional modeling approaches:

1. SEMs implicitly assume that the relationships among variables (paths) are causal. This is a big leap from the traditional statistics' "correlation does not imply causation". By using pre-existing knowledge of the system, one makes an informed hypothesis about the causal structure of the variables, and the SEM explicitly tests this supposed causal structure.
2. Variables can be both predictors and responses. A SEM is thereby useful for testing and quantifying indirect (cascading) effects that would otherwise go unrecognized by any single model.

Traditional SEM

Estimation: Coefficients are estimated simultaneously in a single variance-covariance matrix of all variables; typically by MLE.

Goodness-of-fit: = discrepancy between the observed and predicted covariance matrices. χ_2 test: the χ_2 statistic describes the agreement between the 2 matrices.

Assumptions

- Independent errors (no underlying structure)
- Normal errors

Limits

- *Assumptions often violated in ecological research: e not independent (spatial or temporal correlation in observational studies), distribution not normal (count data ~ Poisson)...*
- computationally intensive (depending on the sizes of the variance-covariance matrix);
- if variables are nested, then the sample size is limited to the use of variables at the highest level of the hierarchy. Can shrink our sample and reduce the power of the analysis...

Piecewise SEM

Estimation: Decompose the network and estimate each relationship separately (estimate m separate vcov matrices). Then piece the m paths together for inferences about the entire SEM.

$\Longrightarrow$ Much easier to estimate than a single vcov matrix → can estimate large networks

$\Longrightarrow$ Flexible: can incorporate many model structures, distributions... using extensions of linear reg (random effects, hierarchical models, non-normal responses, spatial correlation...)

Goodness-of-fit: No formal χ^2 test. Instead: "tests of directed separation": are any paths missing from the model?
The 'basis set' = all k pair relationships unspecified in the model (i.e., independence claims). Test whether are indeed not significant (controlling for variables on which these paths are conditional), keep the p-value. From the k p-values, calculate Fisher's C statistic $C = -2\sum_{i=1}^{k} \ln(p_i) \sim \chi^2(2k)$. If C's p-value > 0.05, accept the model. This approach is vulnerable to model misspecification.
Rmk: we can compute an AIC score for the SEM, for model comparisons: $AIC = C + 2k\frac{n}{n-k-1}$

6.3 Structural Vector Autoregression (SVAR)

Add Ghanem and Smith (2021)

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, DOI: 10.1016/S0304-4076(02)00201-4.

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The Costs of Low Birth Weight. *Q. J. Econ.*, 120(3):1031–1083, DOI: 10.1093/qje/120.3.1031.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 11(1):727–753, DOI: 10.1146/annurev-economics-080218-025643.

Angrist, J. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ, ISBN: 9781400829828, DOI: 10.1515/9781400829828.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier, DOI: 10.1016/bs.hefe.2016.10.003.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in Difference-In-Differences settings with staggered adoption. *J. Econom.*, 226(1):62–79, ISSN: 0304-4076, DOI: 10.1016/j.jeconom.2020.10.012.

Barrios, T., Diamond, R., Imbens, G. W., and Kolesár, M. (2012). Clustering, Spatial Correlations, and Randomization Inference. *J. Am. Stat. Assoc.*, 107(498):578–591, ISSN: 0162-1459, DOI: 10.1080/01621459.2012.682524.

Cooperman, A. D. (2017). Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation. *Polit. Anal.*, 25(3):277–288, DOI: 10.1017/pan.2017.17.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *Am. Econ. Rev.*, 110(9):2964–2996, DOI: 10.1257/aer.20181169.

de Chaisemartin, C. and D'Haultfœuille, X. (2022). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey, DOI: 10.3386/w29691, http://www.nber.org/papers/w29691. Working Paper 29691, National Bureau of Economic Research.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.*, 210:2–21, DOI: 10.1016/j.socscimed.2017.12.005.

Fisher, S. R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Gelman, A. (2011). Causality and Statistical Learning. *Am. J. Sociol.*, 117(3):955–966, DOI: 10.1086/662659.

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press, ISBN: 978-1-107-02398-7, DOI: 10.1017/9781139161879.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456, DOI: 10.1080/07350015.2017.1366909.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: design, analysis, and interpretation*. W. W. Norton & Company, 500 Fifth Avenue, New York, NY 10110-0017, first edition, ISBN: 9780393979954.

Ghanem, D. and Smith, A. (2021). Causality in structural vector autoregressions: Science or sorcery? *Am. J. Agric. Econ.*, DOI: 10.1111/ajae.12269.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *J. Econom.*, 225(2):254–277, ISSN: 0304-4076, DOI: 10.1016/j.jeconom.2021.03.014.

Hayes, R. J. and Moulton, L. H. (2017). *Cluster randomised trials, second edition*. CRC Press, United States, ISBN: 9781498728225, DOI: 10.4324/9781315370286.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *J. Comput. Graph. Stat.*, 20(1):217–240, DOI: 10.1198/jcgs.2010.08162.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, DOI: 10.1093/pan/mpl013.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *J. Econ. Lit.*, 58(4):1129–1179, DOI: 10.1257/jel.20191597.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *J. Econ. Lit.*, 47(1):5–86, ISSN: 0022-0515, DOI: 10.1257/jel.47.1.5.

Kowalski, A. E. (2021). Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform. *The Review of Economics and Statistics*, pages 1–45, DOI: 10.1162/rest_a_01069.

Linden, L. and Rockoff, J. E. (2008). Estimates of the impact of crime risk on property values from megan's laws. *American Economic Review*, 98(3):1103–27, DOI: 10.1257/aer.98.3.1103.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press, ISBN: 9781107065079.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Translated and edited by D.M. Dabrowska and T.P. Speed, Statistical Science (1990), 5, 465-480). *Sci. Ann. Univ. Agric. Sci. Vet. Med.*, 10:1–51, ISSN: 1454-7376.

Pearl, J. (2009). *Causality: models, reasoning, and inference*. Cambridge University Press, New York, second edition, ISBN: 9780521895606.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer series in statistics. Springer Science & Business Media, second edition, ISBN: 9781441931917, DOI: 10.1007/978-1-4757-3692-2.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688–701, DOI: 10.1037/h0037350.

Steiner, P. M., Kim, Y., Hall, C. E., and Su, D. (2017). Graphical Models for Quasi-experimental Designs. *Sociol. Methods Res.*, 46(2):155–188, DOI: 10.1177/0049124115582272.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.*, 225(2):175–199, DOI: 10.1016/j.jeconom.2020.09.006.

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators, DOI: 10.2139/ssrn.3906345, http://dx.doi.org/10.2139/ssrn.3906345. Working Paper.

A Maths of potential outcomes

The steps overlooked in the main document are provided here in blue.

2.1.1 The original selection bias problem

$$\begin{aligned}\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0] &= \mathbb{E}[Y_i^1|D_i{=}1] - \mathbb{E}[Y_i^0|D_i{=}0] \qquad \text{(definition of potential outcomes)}\\ &= \mathbb{E}[Y_i^1|D_i{=}1] - \mathbb{E}[Y_i^0|D_i{=}1] + \mathbb{E}[Y_i^0|D_i{=}1] - \mathbb{E}[Y_i^0|D_i{=}0]\\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid D_i{=}1]}_{\text{ATET}} + \underbrace{\mathbb{E}[Y_i^0 \mid D_i{=}1] - \mathbb{E}[Y_i^0 \mid D_i{=}0]}_{\text{selection bias}}\end{aligned}$$

The same demonstration holds conditional on X_i, i.e., within each stratum of X_i:

$$\begin{aligned}\mathbb{E}[Y_i|D_i{=}1, X_i] - \mathbb{E}[Y_i|D_i{=}0, X_i] &= \mathbb{E}[Y_i^1|D_i{=}1, X_i] - \mathbb{E}[Y_i^0|D_i{=}0, X_i]\\ &= \mathbb{E}[Y_i^1|D_i{=}1, X_i] - \mathbb{E}[Y_i^0|D_i{=}1, X_i] + \mathbb{E}[Y_i^0|D_i{=}1, X_i] - \mathbb{E}[Y_i^0|D_i{=}0, X_i]\\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid D_i{=}1, X_i]}_{\text{ATET for given } X_i} + \underbrace{\mathbb{E}[Y_i^0 \mid D_i{=}1, X_i] - \mathbb{E}[Y_i^0 \mid D_i{=}0, X_i]}_{\text{selection bias for given } X_i}\end{aligned}$$

2.1.2 Expressing TE as a linear regression

Simplest setting: unlimited Y, binary D, no X

The treatment effect can be assumed to be homogeneous or heterogeneous. In either case, we'll show that the linear regression on the treatment recovers the/a treatment effect. The relation between observed outcomes and potential outcomes can be written as a linear regression on the treatment:

- Case 1: homogeneous treatment effect $Y_i^1 - Y_i^0 = \beta$

$$\begin{aligned}Y_i &= Y_i^0 + \left(Y_i^1 - Y_i^0\right) D_i\\ &= \mathbb{E}[Y_i^0] + \beta D_i + Y_i^0 - \mathbb{E}[Y_i^0]\\ &= \alpha + \beta D_i + u_i\end{aligned}$$

- Case 2: heterogeneous treatment effect $Y_i^1 - Y_i^0 = \beta_i$. Note β the ATET $\mathbb{E}[\beta_i \mid D_i{=}1]$.

$$\begin{aligned}Y_i &= Y_i^0 + \left(Y_i^1 - Y_i^0\right) D_i\\ &= \mathbb{E}[Y_i^0] + \beta_i D_i + Y_i^0 - \mathbb{E}[Y_i^0]\\ &= \mathbb{E}[Y_i^0] + \beta D_i + (\beta_i - \beta) D_i + Y_i^0 - \mathbb{E}[Y_i^0]\\ &= \alpha + \beta D_i + u_i\end{aligned}$$

The OLS slope estimand simplifies to the difference in average observed outcomes, which itself simplifies to an expression with the error term:

$$\begin{aligned}
\beta_{\text{OLS}} &= \frac{\text{cov}[Y_i, D_i]}{\mathbb{V}[D_i]} = \frac{\mathbb{E}[Y_i D_i] - \mathbb{E}[Y_i]\mathbb{E}[D_i]}{\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2} \\
&= \frac{\mathbb{E}[Y_i|D_i{=}1]\ \text{P}(D_i{=}1) - \Big(\mathbb{E}[Y_i|D_i{=}0]\ \text{P}(D_i{=}0) + \mathbb{E}[Y_i|D_i{=}1]\ \text{P}(D_i{=}1)\Big)\ \text{P}(D_i{=}1)}{\text{P}(D_i{=}1) - \text{P}(D_i{=}1)^2} \\
&= \frac{\mathbb{E}[Y_i|D_i{=}1]\ \text{P}(D_i{=}1)\ \big(1 - \text{P}(D_i{=}1)\big) - \mathbb{E}[Y_i|D_i{=}0]\ \text{P}(D_i{=}0)\ \text{P}(D_i{=}1)}{\text{P}(D_i{=}1)\big(1 - \text{P}(D_i{=}1)\big)} \\
&= \frac{\mathbb{E}[Y_i|D_i{=}1]\ \text{P}(D_i{=}1)\ \text{P}(D_i{=}0) - \mathbb{E}[Y_i|D_i{=}0]\ \text{P}(D_i{=}0)\ \text{P}(D_i{=}1)}{\text{P}(D_i{=}1)\ \text{P}(D_i{=}0)} \\
&= \mathbb{E}[Y_i \mid D_i{=}1] - \mathbb{E}[Y_i \mid D_i{=}0]
\end{aligned}$$

$$\begin{cases} \mathbb{E}[Y_i|D_i{=}1] = \alpha + \beta + \mathbb{E}[u_i|D_i{=}1] \\ \mathbb{E}[Y_i|D_i{=}0] = \alpha + \mathbb{E}[u_i|D_i{=}0] \end{cases} \implies \mathbb{E}[Y_i \mid D_i{=}1] - \mathbb{E}[Y_i \mid D_i{=}0] \;=\; \beta + \mathbb{E}[u_i \mid D_i{=}1] - \mathbb{E}[u_i \mid D_i{=}0]$$

- Case 1: $u_i \equiv Y_i^0 - \mathbb{E}[Y_i^0]$, therefore:

$$\mathbb{E}[u_i \mid D_i{=}1] - \mathbb{E}[u_i \mid D_i{=}0] = \mathbb{E}[Y_i^0 \mid D_i{=}1] - \mathbb{E}[Y_i^0 \mid D_i{=}0]$$

- Case 2: $u_i \equiv (\beta_i - \beta)D_i + Y_i^0 - \mathbb{E}[Y_i^0]$, therefore:

$$\begin{aligned}
\mathbb{E}[u_i \mid D_i{=}1] - \mathbb{E}[u_i \mid D_i{=}0] &= \mathbb{E}[\beta_i{-}\beta \mid D_i{=}1] + \mathbb{E}[Y_i^0 \mid D_i{=}1] - 0 - \mathbb{E}[Y_i^0] - \mathbb{E}[Y_i^0 \mid D_i{=}0] + \mathbb{E}[Y_i^0] \\
&= \mathbb{E}[\beta_i \mid D_i{=}1] - \beta + \mathbb{E}[Y_i^0 \mid D_i{=}1] - \mathbb{E}[Y_i^0 \mid D_i{=}0] \\
&= \mathbb{E}[Y_i^0 \mid D_i{=}1] - \mathbb{E}[Y_i^0 \mid D_i{=}0]
\end{aligned}$$

In both cases, $\beta_{\text{OLS}} = \ldots = \mathbb{E}[Y_i \mid D_i{=}1] - \mathbb{E}[Y_i \mid D_i{=}0] = \ldots = \beta + \text{selection bias}$.

With covariates X

For simplicity, consider a discrete X_i.

- Matching estimand

$$\begin{aligned}
\beta_{\text{M}} &= \sum_x \delta_x\ \text{P}(X_i{=}x|D_i{=}1) = \sum_x \delta_x \frac{\text{P}(X_i{=}x, D_i{=}1)}{\text{P}(D_i{=}1)} = \sum_x \delta_x \frac{\text{P}(D_i{=}1|X_i{=}x)\ \text{P}(X_i{=}x)}{\text{P}(D_i{=}1)} \\
&= \frac{1}{\text{P}(D_i{=}1)} \sum_x \delta_x\ \text{P}(D_i{=}1|X_i{=}x)\text{P}(X_i{=}x) \\
&= \frac{\sum_x \delta_x\ \text{P}(D_i{=}1|X_i{=}x)\text{P}(X_i{=}x)}{\sum_x \text{P}(D_i{=}1|X_i{=}x)\text{P}(X_i{=}x)}
\end{aligned}$$

- OLS estimand
  The demonstration uses the Frisch-Waugh-Lovell theorem (Angrist and Pischke, 2008, p.55).

3.1 IV

IV estimand

$$\beta_{\text{IV}} \equiv \frac{\text{cov}[Y_i, Z_i]}{\text{cov}[D_i, Z_i]} = \frac{\mathbb{E}[Y_i Z_i] - \mathbb{E}[Y_i]\mathbb{E}[Z_i]}{\mathbb{E}[D_i Z_i] - \mathbb{E}[D_i]\mathbb{E}[Z_i]}$$

$$= \frac{\mathbb{E}[Y_i \mid Z_i{=}1]\text{P}(Z_i{=}1) - \Big(\mathbb{E}[Y_i \mid Z_i{=}1]\text{P}(Z_i{=}1) + \mathbb{E}[Y_i \mid Z_i{=}0]\text{P}(Z_i{=}0)\Big)\text{P}(Z_i{=}1)}{\mathbb{E}[D_i \mid Z_i{=}1]\text{P}(Z_i{=}1) - \Big(\mathbb{E}[D_i \mid Z_i{=}1]\text{P}(Z_i{=}1) + \mathbb{E}[D_i \mid Z_i{=}0]\text{P}(Z_i{=}0)\Big)\text{P}(Z_i{=}1)}$$

$$= \frac{\mathbb{E}[Y_i \mid Z_i{=}1]\big(1 - \text{P}(Z_i{=}1)\big) - \mathbb{E}[Y_i \mid Z_i{=}0]\text{P}(Z_i{=}0)}{\mathbb{E}[D_i \mid Z_i{=}1]\big(1 - \text{P}(Z_i{=}1)\big) - \mathbb{E}[D_i \mid Z_i{=}0]\text{P}(Z_i{=}0)}$$

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

The identifying assumptions then reduce it to the LATE on the compliers:

- Numerator: $\mathbb{E}\big[Y_i \mid Z_i{=}1\big] - \mathbb{E}\big[Y_i \mid Z_i{=}0\big] =$

  $= \mathbb{E}\big[Y_i|Z_i{=}1, D_i^0{=}0, D_i^1{=}0\big]P(D_i^0{=}0, D_i^1{=}0) - \mathbb{E}\big[Y_i|Z_i{=}0, D_i^0{=}0, D_i^1{=}0\big]P(D_i^0{=}0, D_i^1{=}0)$
  $+ \mathbb{E}\big[Y_i|Z_i{=}1, \text{—}\,0, \text{—}\,1\big]P(\text{—}\,0, \text{—}\,1) - \mathbb{E}\big[Y_i|Z_i{=}0, \text{—}\,0, \text{—}\,1\big]P(\text{—}\,0, \text{—}\,1)$
  $+ \mathbb{E}\big[Y_i|Z_i{=}1, \text{—}\,1, \text{—}\,0\big]P(\text{—}\,1, \text{—}\,0) - \mathbb{E}\big[Y_i|Z_i{=}0, \text{—}\,1, \text{—}\,0\big]P(\text{—}\,1, \text{—}\,0)$
  $+ \mathbb{E}\big[Y_i|Z_i{=}1, \text{—}\,1, \text{—}\,1\big]P(\text{—}\,1, \text{—}\,1) - \mathbb{E}\big[Y_i|Z_i{=}0, \text{—}\,1, \text{—}\,1\big]P(\text{—}\,1, \text{—}\,1)$
  $= \cancel{\mathbb{E}\big[Y_i^0|D_i^0{=}0, D_i^1{=}0\big]}P(D_i^0{=}0, D_i^1{=}0) - \cancel{\mathbb{E}\big[Y_i^0|D_i^0{=}0, D_i^1{=}0\big]}P(D_i^0{=}0, D_i^1{=}0)$
  $+ \mathbb{E}\big[Y_i^1|\text{—}\,0, \text{—}\,1\big]P(\text{—}\,0, \text{—}\,1) - \mathbb{E}\big[Y_i^0|\text{—}\,0, \text{—}\,1\big]P(\text{—}\,0, \text{—}\,1)$
  $+ \mathbb{E}\big[Y_i^0|\text{—}\,1, \text{—}\,0\big]P(\text{—}\,1, \text{—}\,0) - \mathbb{E}\big[Y_i^1|\text{—}\,1, \text{—}\,0\big]P(\text{—}\,1, \text{—}\,0)$
  $+ \cancel{\mathbb{E}\big[Y_i^1|\text{—}\,1, \text{—}\,1\big]}P(\text{—}\,1, \text{—}\,1) - \cancel{\mathbb{E}\big[Y_i^1|\text{—}\,1, \text{—}\,1\big]}P(\text{—}\,1, \text{—}\,1)$
  $= \mathbb{E}\big[Y_i^1 - Y_i^0 \mid D_i^0{=}0, D_i^1{=}1\big]P(D_i^0{=}0, D_i^1{=}1) - \mathbb{E}\big[Y_i^1 - Y_i^0 \mid D_i^0{=}1, D_i^1{=}0\big]P(D_i^0{=}1, D_i^1{=}0)$
  $= \mathbb{E}\big[Y_i^1 - Y_i^0 \mid D_i^0{=}0, D_i^1{=}1\big]P(D_i^0{=}0, D_i^1{=}1)$ as the probability of defiers is 0

- Denominator:
  $\mathbb{E}\big[D_i \mid Z_i{=}1\big] - \mathbb{E}\big[D_i \mid Z_i{=}0\big] = \mathbb{E}\big[D_i^1 - D_i^0\big]$
  $= 1 \times P\big(D_i^1 - D_i^0 = 1\big) + 0 \times P\big(D_i^1 - D_i^0 = 0\big) - 1 \times P\big(D_i^1 - D_i^0 = -1\big)$
  $= P\big(D_i^0{=}0, D_i^1{=}1\big)$ as the probability of defiers is 0

$$\Longrightarrow \frac{\mathbb{E}[Y_i \mid Z_i{=}1] - \mathbb{E}[Y_i \mid Z_i{=}0]}{\mathbb{E}[D_i \mid Z_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0]} = \underbrace{\mathbb{E}\big[Y_i^1 - Y_i^0 \mid D_i^0{=}0, D_i^1{=}1\big]}_{\textbf{LATE on the compliers}}$$

Counting and characterizing compliers to get more out of the LATE

- Size of the complier group: It is the Wald first-stage, as, given monotonicity:

$$P[D_i^1 > D_i^0] = P[D_i^1 - D_i^0 = 1] = \mathbb{E}[D_i^1 - D_i^0] = \mathbb{E}[D_i^1] - \mathbb{E}[D_i^0] = \mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0]$$

- Share of treated that are compliers:

$$\begin{aligned} P[D_i^1 > D_i^0 \mid D_i{=}1] &= \frac{P[D_i^1 > D_i^0, D_i{=}1]}{P[D_i{=}1]} = \frac{P[D_i{=}1 \mid D_i^1 > D_i^0] \;\; P[D_i^1 > D_i^0]}{P[D_i{=}1]} \\ &= \frac{P[Z_i{=}1 \mid D_i^1 > D_i^0] \;\; P[D_i^1 > D_i^0]}{P[D_i{=}1]} \\ &= \frac{P[Z_i{=}1] \;\; P[D_i^1 > D_i^0]}{P[D_i{=}1]} \\ &= \frac{P[Z_i{=}1] \times (\mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0])}{P[D_i{=}1]} \\ &= \frac{P[\text{instrument is switched on}] \times \text{1st stage}}{\text{share treated}} \end{aligned}$$

- Distribution of covariates X for compliers:
  - For binary characteristics X, we can compute relative likelihoods:

$$\begin{aligned} \frac{P[X_i{=}1 \mid D_i^1 > D_i^0]}{P[X_i{=}1]} &= \frac{P[X_i{=}1, D_i^1 > D_i^0]}{P[X_i{=}1] \;\; P[D_i^1 > D_i^0]} = \frac{P[D_i^1 > D_i^0 \mid X_i{=}1]}{P[D_i^1 > D_i^0]} \\ &= \frac{\mathbb{E}[D_i \mid Z_i{=}1, X_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0, X_i{=}1]}{\mathbb{E}[D_i \mid Z_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0]} \\ &= \frac{\text{1st stage} \mid X_i{=}1}{\text{1st stage}} \end{aligned}$$

3.2 RD

Sharp RD estimand

$$\begin{aligned} \beta_{\text{RD}} &\equiv \lim_{x \to c^+} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \to c^-} \mathbb{E}[Y_i \mid X_i = x] \\ &= \lim_{x \to c^+} \mathbb{E}[Y_i^1 \mid X_i = x] - \lim_{x \to c^-} \mathbb{E}[Y_i^0 \mid X_i = x] \\ &= \mathbb{E}[Y_i^1 \mid X_i = c] - \mathbb{E}[Y_i^0 \mid X_i = c] \\ &= \underbrace{\mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = c]}_{\textbf{LATE at the cutoff}} \end{aligned}$$

Fuzzy RD estimand

$$\beta_{\text{IV}} \equiv \frac{\lim_{x \to c^+} \mathbb{E}[Y_i \mid X_i{=}x] - \lim_{x \to c^-} \mathbb{E}[Y_i \mid X_i{=}x]}{\lim_{x \to c^+} \mathbb{E}[D_i \mid X_i{=}x] - \lim_{x \to c^-} \mathbb{E}[D_i \mid X_i{=}x]} = \lim_{\delta \to 0} \frac{\mathbb{E}[Y_i \mid c < X_i < c + \delta] - \mathbb{E}[Y_i \mid c - \delta < X_i < c]}{\mathbb{E}[D_i \mid c < X_i < c + \delta] - \mathbb{E}[D_i \mid c - \delta < X_i < c]}$$

$\mathbb{E}[Y_i \mid c < X_i < c + \delta] - \mathbb{E}[Y_i \mid c - \delta < X_i < c] \simeq \gamma\beta$
$\mathbb{E}[D_i \mid c < X_i < c + \delta] - \mathbb{E}[D_i \mid c - \delta < X_i < c] \simeq \gamma$

Therefore, $\beta_{\text{IV}} = \dfrac{\gamma\beta}{\gamma} = \beta$

3.3 DiD, DiDiD, Event-study

DiD estimand

$$
\begin{aligned}
\beta_{\text{DiD}} &\equiv \left(\bar{Y}_{G_1P_1} - \bar{Y}_{G_1P_0}\right) - \left(\bar{Y}_{G_0P_1} - \bar{Y}_{G_0P_0}\right) \\
&\equiv \Big(\mathbb{E}\big[Y_{i1} \mid G_i{=}1\big] - \mathbb{E}\big[Y_{i0} \mid G_i{=}1\big]\Big) - \Big(\mathbb{E}\big[Y_{i1} \mid G_i{=}0\big] - \mathbb{E}\big[Y_{i0} \mid G_i{=}0\big]\Big) \\
&= \Big(\mathbb{E}\left[Y^1_{i1} \mid G_i{=}1\right] - \mathbb{E}\left[Y_{i0} \mid G_i{=}1\right]\Big) - \Big(\mathbb{E}\left[Y^0_{i1} \mid G_i{=}0\right] - \mathbb{E}\left[Y_{i0} \mid G_i{=}0\right]\Big) \\
&= \mathbb{E}\left[Y^1_{i1} - Y_{i0} \mid G_i{=}1\right] - \mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_i{=}0\right] \\
&= \mathbb{E}\left[Y^1_{i1} - Y_{i0} \mid G_i{=}1\right] - \mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_i{=}1\right] \qquad \text{(assumption of parallel trends)} \\
&= \underbrace{\mathbb{E}\big[Y^1_{i1} - Y^0_{i1} \mid G_i{=}1\big]}_{\textbf{ATET}}
\end{aligned}
$$

DiDiD estimand

$$
\begin{aligned}
\beta_{\text{DiDiD}} &\equiv \Big[\left(\bar{Y}_{G_1S_1P_1} - \bar{Y}_{G_1S_1P_0}\right) - \left(\bar{Y}_{G_1S_0P_1} - \bar{Y}_{G_1S_0P_0}\right)\Big] - \Big[\left(\bar{Y}_{G_0S_1P_1} - \bar{Y}_{G_0S_1P_0}\right) - \left(\bar{Y}_{G_0S_0P_1} - \bar{Y}_{G_0S_0P_0}\right)\Big] \\
&= \Big(\mathbb{E}\left[Y_{i1} - Y_{i0} \mid G_1, S_1\right] - \mathbb{E}\left[Y_{i1} - Y_{i0} \mid G_1, S_0\right]\Big) - \Big(\mathbb{E}\left[Y_{i1} - Y_{i0} \mid G_0, S_1\right] - \mathbb{E}\left[Y_{i1} - Y_{i0} \mid G_0, S_0\right]\Big) \\
&= \Big(\mathbb{E}\left[Y^1_{i1} - Y_{i0} \mid G_1, S_1\right] - \mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_1, S_0\right]\Big) - \Big(\mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_0, S_1\right] - \mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_0, S_0\right]\Big) \\
&= \Big(\mathbb{E}\left[Y^1_{i1} - Y_{i0} \mid G_1, S_1\right] - \cancel{\mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_1, S_0\right]}\Big) - \Big(\mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid \mathbf{G_1}, \mathbf{S_1}\right] - \cancel{\mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid \mathbf{G_1}, \mathbf{S_0}\right]}\Big) \quad (\parallel \text{ trends}) \\
&= \mathbb{E}\left[Y^1_{i1} - Y_{i0} \mid G_1, S_1\right] - \mathbb{E}\left[Y^0_{i1} - Y_{i0} \mid G_1, S_1\right] \\
&= \underbrace{\mathbb{E}\left[Y^1_{i1} - Y^0_{i1} \mid G_1, S_1\right]}_{\textbf{ATET}}
\end{aligned}
$$