Causal Inference in Observational Studies

Contents

Disclaimers:

- *Sections and lines in brown correspond to content which is* ***very much*** *'under construction'.*
- *For all expressions whose simplification into a final expression is not detailed (either explicitly stated or by using "..."), the mathematical steps of the simplification 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 Theory: 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 treatment of interest $D_i \in \{0,1\}$ whose causal effect we want to estimate. Let Y_i be the realized outcome, Y_i^0, Y_i^1 the potential outcomes. 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*
– Difference in average observed outcomes for same X_i	$\mathbb{E}[Y_i \vert D_i{=}1, X_i] - \mathbb{E}[Y_i \vert D_i{=}0, X_i]$	*what we can compute*

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

- To measure TE $= Y_i^1 - Y_i^0$,[4] we need to observe the same individual with and without treatment.
- This is impossible, we can't observe the counterfactual.[5] We can only compute the difference in average observed outcomes:

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

  The "selection bias" is the average difference in Y_i^0 between the treated and untreated.[6]
- 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 over all trials*. The independence assumption (IA) identifies the ATET.
- 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 the potential outcomes are *conditionally* independent of the treatment $(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i | X_i$, where X is a vector of confounding pre-treatment characteristics, then we again eliminate selection bias. The conditional independence assumption (CIA) identifies the ATET.

We can estimate an *unbiased* causal effect iff an **identifying/independence[7] assumption** holds:

– if IA	$(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i$	$\Longrightarrow$	we can estimate the ATET.
– if ~~IA~~ but CIA	$(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i \vert X_i$	$\Longrightarrow$	we can estimate the ATET.
– if ~~CIA~~ but $\exists$ a relevant instrument Z that is an exogenous source of variation in D:	$(Y_i^0, Y_i^1) \perp\!\!\!\perp D_i, \quad Z_i \not\perp\!\!\!\perp D_i \vert X_i$	$\Longrightarrow$	we can estimate *a* LATE.

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

[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] The convention in the causal inference literature is to define a causal effect as a linear function of the potential outcomes, i.e., as $Y_i^1 - Y_i^0$. Note that causal effects could also be expressed as nonlinear functions, e.g., as Y_i^1 / Y_i^0.

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

[6] 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.

[7] 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 treatment assignment process to estimate causal effects, we need only compare group means. Examples of assignment mechanism: random assignment (the IA); selection on observables (the CIA); selection on unobservables...

Note: The identification result extends beyond average treatment effects. Independence identifies the entire marginal distributions of the potential outcomes, and therefore also quantile treatment effects — i.e, $\forall p$, the effect of the treatment at quantile p: $\tau_p \equiv \mathbb{Q}_{Y^1}(p) - \mathbb{Q}_{Y^0}(p)$.[8] One may be interested in TEs that may be concentrated in tails of the distribution of outcomes, i.e., treatments that affect a tail more than the center of the distribution, or in estimating more robustly constant TEs 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 $\beta_{\text{OLS}} \equiv \frac{cov[Y_i, D_i]}{Var[D_i]}$ simplifies to $\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0]$: the difference in average observed outcomes.
- This, given the regression equation, equals $\beta \; + \; \mathbb{E}[u_i|D_i{=}1] - \mathbb{E}[u_i|D_i{=}0]$.
- If u_i is uncorrelated with D_i, then $\beta \; + \; \mathbb{E}[u_i|D_i{=}1] - \mathbb{E}[u_i|D_i{=}0] = ...$ = ATET + 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).[9]

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

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., $cov[D_i, e_i] \neq 0$, 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;

[8] ⚠ The p-th QTE $\mathbb{Q}_{Y^1}(p) - \mathbb{Q}_{Y^0}(p)$ is the effect of the treatment at quantile p, not the p–th quantile of the treatment effect $\mathbb{Q}_{Y^0 - Y^1}(p)$. A difference in quantiles is generally not equal to the quantile of differences.

[9] ⚠ With many covariates, the nonparametric analysis by computing the difference in averages becomes too complicated, so we are even more incentivized to use regression. Note however that one should expect differences between the matching estimator and the regression estimator (Angrist and Pischke, 2008, 3.3.1). The two estimands differ in the weights used to sum the estimates of the covariate-specific effects $\delta_X \equiv \mathbb{E}[Y_i|X_i, D_i{=}1] - \mathbb{E}[Y_i|X_i, D_i{=}0]$ into a single ATET. In the OLS estimand, the weights are proportional to the conditional variance of $D_i|X_i$ — which is maximized when $P(D_i{=}1|X_i{=}x) = .5$, i.e., for values of X_i with as many treated as control observations. OLS is a minimum-variance estimator: it gives more weight to more precise within-strata estimates. Whereas in the matching estimand, the weights are proportional to the conditional probability of treatment $P(D_i{=}1|X_i{=}x)$. This difference in weighting schemes induces 2 different estimates only if δ_X varies along X, i.e., only if the TE is heterogeneous. Then we may want to interact D with X.

- omitted variable bias (OVB): All omitted variables[10] are captured by e. Therefore, if an omitted variable w is correlated with D, e is correlated with D. w is a "confounding variable".

> → In observational studies,
> – excluding a confounding variable creates bias, so we must adjust for all *confounders*.
> – with all confounders assumed to be measured, we estimate an unbiased causal effect.
> – because we can rarely be certain to have measured all confounders,[11] we turn to alternative causal inference or **"identification" strategies**, that rely on other assumptions.

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

[11] For instance, in cross-sectional approaches, we worry about time-invariant omitted variables. As a cross-section offers only inter-individual (across) variation, if y is affected by unobservable variables that systematically vary across groups, our estimate will be biased. With panel data, we have across variation and intra-individual (within) variation. Using individual fixed effects, we can focus on within variation only, 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.[12] 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 causal states differently, DAGs are just a language to make the identifying assumptions explicit.

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.[13]

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:

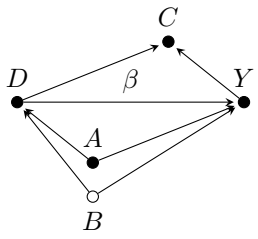
- a node represents a random variable; a solid circle if it is observed, hollow otherwise;
- all edges are directed and represent causal relationships;
- a path is any sequence of edges;
- a **back-door path** = any path between D and Y that begins with an arrow pointing to D.[a] It is *closed* if at 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.



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

[13] 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.

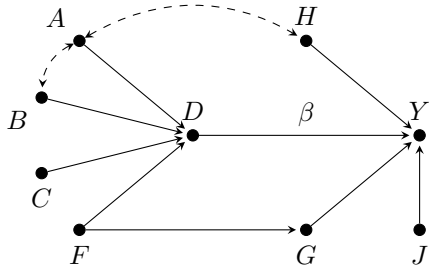
2.2.2 Two identification strategies: 1. blocking back-door paths; 2. instruments

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 path.
    - However the back-door path through B is *open*, as B is unobserved. → We therefore cannot recover β by blocking back-door paths.
  - In the more complex DAG on the right, there are three back-door paths:[14]

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

    We can block all back-door paths by either:
    - conditioning on H and either F or G
    - conditioning on A and B,[15] 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[16] 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 measured and conditioned on, we estimate an unbiased causal effect.
> – because we can rarely be certain that we have measured all confounders, we turn to alternative causal inference or **"identification" strategies**, that rely on other assumptions.

[14] 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.

[15] 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.

[16] 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

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 (Gelman, 2011).

3 Design stage: Applied identification methods

3.1 Hierarchy of common identification methods

A contestable hierarchy of the most common identification methods in the 'randomista' toolkit,[17] based on their capacity to mimic random assignment, is as follows:

0. Randomized experiment (RCT) — or direct natural randomization of treatment D
1. Instrumental Variables (IV) and regression discontinuity (RD)
   If we fear that there is 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 rely on the assumptions of time-invariant omitted variables and parallel trends.
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 underlies how our estimation recovers a causal treatment effect.

For simplification purposes, all methods are presented without the inclusion of exogenous controls X_i. However, the relationships hold when they are all conditional on covariates X_i.

[17] Term shamelessly copied from Gibson (2019).

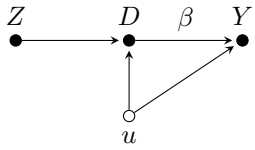
3.2 Random assignment

Show:

- experimental: RCT
- observational but still random assignment: natural experiment. Ex: weather or other

3.3 IV

Data Generating Process (DGP) $Y_i = \alpha + \beta_i\, D_i + u_i, \quad cov[D_i, u_i] \neq 0$: 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:[18]



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

In terms of 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, which will be 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: $cov[Z_i, u_i] = 0$ *(and* $cov[Z_i, covariates_i] = 0$*)*. I.e., Z_i affects Y_i only through D_i

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

(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 We define the IV estimand: $\beta_{\text{IV}} \equiv \dfrac{cov[Y_i, Z_i]}{cov[D_i, Z_i]} = ... = \dfrac{\mathbb{E}\big[Y_i \mid Z_i{=}1\big] - \mathbb{E}\big[Y_i \mid Z_i{=}0\big]}{\mathbb{E}\big[D_i \mid Z_i{=}1\big] - \mathbb{E}\big[D_i \mid Z_i{=}0\big]}$. Note that:

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

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

$\implies$ Their ratio $\widehat{\beta}_{\text{IV}} \equiv \frac{\widehat{\gamma.\beta}_{\text{LS}}}{\widehat{\gamma}_{\text{LS}}} = ... = \beta + \frac{\widehat{cov}[u_i, Z_i]}{\widehat{cov}[D_i, Z_i]}$: is consistent but has a bias, which $\searrow$ with Z_i's strength.

The identifying assumptions reduce $\dfrac{\mathbb{E}\,[Y_i \mid Z_i{=}1] - \mathbb{E}\,[Y_i \mid Z_i{=}0]}{\mathbb{E}\,[D_i \mid Z_i{=}1] - \mathbb{E}\,[D_i \mid Z_i{=}0]}$ to $\underbrace{\mathbb{E}\big[Y_i^1 - Y_i^0 \mid D_i^0{=}0, D_i^1{=}1\big]}_{\textbf{LATE on the compliers}}$

Estimator Our natural choice of estimator is the sample analog called "Wald estimator" $\widehat{\beta}_{\text{W}} = \frac{\widehat{cov}[Y_i, Z_i]}{\widehat{cov}[D_i, Z_i]}$. It turns out to be numerically equivalent to the 2SLS estimator $\widehat{\beta}_{\text{2SLS}}$ obtained through the two-step procedure:[19]

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

[18] 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.

[19] 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. Or run a test such as the Stock and Yogo test.
- As in any observational study, adjust for **all other** relevant pre-treatment variables (include the same variables in both stages).
- Different valid instruments will yield different estimates because they correspond to different estimands, as each selected a specific set of compliers. 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 the compliers to get more out of the LATE.
- For models that are non-linear in D_i, the properties of 2SLS do not necessarily hold. Consider alternative estimation strategies (e.g., 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 generally more efficient but less robust than 2SLS as it imposes additional restrictions.

Strengths & weaknesses

+ Compelling identification strategy
– Strong assumptions
– Beware of weak instruments. The 2SLS estimator is biased, and its bias increases with the weakness and number of instruments. It will also be less efficient than the OLS estimator if instruments are weak. On weak instruments, see Andrews et al. (2019).
– In many settings (e.g., models non-linear in D, non-saturated models with covariates) 2SLS does not fare well, i.e., can be very biased. New paper: `https://a-torgovitsky.github.io/tsls-weights.pdf`

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 $\kappa_i = 1 - \frac{D_i(1-Z_i)}{1-P[Z_i=1|X_i]} - \frac{(1-D_i)Z_i}{P[Z_i=1|X_i]}$

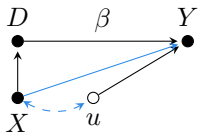
Applications: Kowalski (2021) (Data and code: `https://doi.org/10.7910/DVN/UGMXUQ`)

3.4 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 $\mathbb{E}\left[Y_i^0|X_i\right] = f(X_i)$, the DGP is described below, where the blue arrows disappear as $X \to c$:[20]



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

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

Estimator We can estimate β at the cutoff by running the centered regression below:[21]

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

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 $c - h \leq X \leq c + h$.[22]
  - 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.

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

[21] 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 $f()$: $Y_i = \alpha + \beta.D_i + f(X_i, \phi_l) + f(X_i, \phi_r)D_i + e_i$

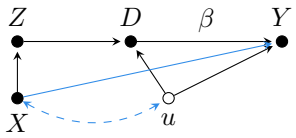
[22] A larger bandwidth h increases precision but also bias. Choose the optimal h by estimating the model's predictive accuracy for different values of h, for example using leave-one-out cross-validation: iteratively for each observation i, fit the model using only the observations $X_i - h \leq \mathbf{X} < X_i < c$ 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 $P[D_i = 1|X_i]$. 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 blue arrows disappear as $X \to c$:



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

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

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

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

3.5 DiD, DiDiD

Identification from variation within & between groups

DiD

DGP Treatment assignment or exposure is a function of 2 dimensions: group (treatment/control) and most commonly time (pre/post exposure).[23] We define the associated binary variables $G_i \equiv \mathbb{1}\{i \in \text{treatment group}\}$ and $P_t \equiv \mathbb{1}\{t \in \text{post period}\}$.

In a before/after comparison among the treatment group, the difference in Y could be the result of other changes that occurred during the time period; in a treatment/control groups comparison in the post period, the difference in Y could be the result of 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. We implicitly adjust for *time-invariant* differences across the groups.

Identifying assumptions

(A1) Same counterfactual trends across groups, i.e., same *potential changes (pre to post) in outcomes* Y_{it} in the absence of treatment: $\mathbb{E}\big[Y^0_{i1} - Y_{i0} \mid G_i{=}1\big] = \mathbb{E}\big[Y^0_{i1} - Y_{i0} \mid G_i{=}0\big]$

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

Estimand

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

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

$$\begin{aligned} Y_{it} &= \alpha + \beta_G G_i + \beta_P P_t + \boldsymbol{\beta}\, G_i P_t + e_{it} \\ &= \lambda_G + \lambda_P + \boldsymbol{\beta}\, G_i P_t + e_{it} \end{aligned}$$

If we have panel data (instead of merely repeated cross-sections), we can estimate this in a more direct way with a first difference approach, by regressing the change for each unit $Y_{i1} - Y_{i0}$ on G_i.

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 regression model by OLS, and check that the coefficients β_τ where $\tau < t_0 - 1$ are 0:

$$y_{it} = \sum_{\tau \neq t_0 - 1} \beta_\tau G_i \mathbb{1}\{t = \tau\} + \lambda_i + \lambda_t + e_{it}$$

- The regression above also enables us to look at whether the TE *accumulates* over time: $\beta_{\tau, \tau \geq t_0} \uparrow$ in τ.
- If the composition of the groups changes over time, interact covariates with P_t.
- As in any observational study, adjust for **all other** relevant pre-treatment variables.

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

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" (in addition to time and group), such as gender, space... We define the binary variable $S_i \equiv \mathbb{1}\{i \in \text{treatment in dim \#3}\}$.

Identifying assumptions

(A1) Same counterfactual trends across ~~groups~~ subgroups, i.e., same changes in outcomes when no treatment:

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

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

Estimand

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

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

$$\begin{aligned}
Y_{it} &= \alpha + \beta_G G_i + \beta_S S_i + \beta_P P_t + \beta_{GS} G_i S_i + \beta_{GP} G_i P_t + \beta_{PS} P_t S_i + \boldsymbol{\beta}\, G_i S_i P_t + e_{it} \\
&= \boldsymbol{\beta}\, G_i S_i P_t + \lambda_{GS} + \lambda_{GP} + \lambda_{PS} + e_{it}
\end{aligned}$$

Best practices

- A triple differences makes for a very specific control group. Before doing an DiDiD, one must be able to answer why a double differences wasn't satisfactory (why the control group in double differences isn't good enough), and even the first differences.
- As in any observational study, adjust for **all other** relevant pre-treatment variables.

Strengths & weaknesses

+ A triple difference allows to difference out more confounding elements, it therefore gets harder to find a confounder.
– It requires more data and variation.

3.6 Event study

DGP We want to estimate the causal effect of *an event*,[24] which may occur at different times τ_i for each unit i ("staggered adoption") and affects *all units* in the population, on some outcome Y. Treatment assignment is a function of the period (pre/post event). We define the binary variable $P_t \equiv \mathbb{1}\{t \in \text{post period}\}$.

Identifying assumptions

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

Estimand

$$\begin{aligned}\beta_{\text{ES}} &\equiv \mathbb{E}\big[Y_{it} \mid t = \tau_i\big] - \mathbb{E}\big[Y_{it} \mid t = \tau_i - 1\big] \\ &= \mathbb{E}\big[Y^1_{i,\tau_i}\big] - \mathbb{E}\left[Y_{i,\tau_i - 1}\right] \\ &= \mathbb{E}\big[Y^1_{i,\tau_i}\big] - \mathbb{E}\left[Y^0_{i,\tau_i}\right] \\ &= \underbrace{\mathbb{E}\big[Y^1_{i\tau_i} - Y^0_{i\tau_i}\big]}_{\textbf{ATET}}\end{aligned}$$

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

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

Best practices

- Report all β_ts, to check that they are not increasing up to the event. An increase would suggest the presence of pre-trends... which are a sign of endogeneity of the treatment variable, making it hard to interpret the event (unless there is a trend discontinuity). Provide a plot of pre-trends.
- 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.

[24] Like in DiD, we are estimating the causal effect of an event, thanks to observing units repeatedly over time. We need a model to estimate the counterfactual value (if the event had not occurred), s.t. the difference from the counterfactual is the causal effect. DiD and event studies are simply different models of the counterfactual. We use DiD when there are control units that we can use to remove trends in the outcome of interest.

3.7 SCM

Summary of common identification methods

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

4 Analysis stage: Steps for stronger causal inferences

4.1 Identification strategies provide only so much

Recall the core motivation for identification strategies:

> We look for identification strategies that suggest that an independence assumption holds, as:
> - if IA *(e.g., in an RCT)*, the regression of Y on D gives an unbiased estimate of the ATET;
> - if ~~IA~~ but 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, then 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 correct estimation of the unbiased ATET relies on correctly specifying the functional form w.r.t. X. But we don't ever know this $f(X)$ for sure. Then, we must avoid by all means areas of imperfect overlap in our data, as in these areas, we are forced to rely on model specification instead of direct support from the data, thus the less overlap in our data, the least robust our inferences are to model misspecification. Assuming all confounders are observed and 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.
2. An unbiased estimator does not guarantee that the estimate from any particular sample/randomization will be close to the true value of the estimand — especially with a small sample size.[25] We might therefore want to:
   (a) adjust as much as possible for potential imbalance between the two groups, by adjusting for pre-treatment information;
   (b) reduce the width of the distribution of the estimator (i.e., consider another property: efficiency).
3. We obtain an estimate of the ATET. What are we learning from it? Reduced forms are generally — this document is no exception — motivated by placing the RCT as gold standard. In the context of an RCT, where the treatment variable represents an intervention, the average effect of that intervention might very well be exactly the knowledge desired. But in other contexts, knowing the magnitude of the relationship but not the mechanisms[26] might be considerably less interesting, e.g., in studying the impact of climate extremes on social instability.

This section goes over what can be done at the analysis stage (i.e., post-design, with a given 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. the observed confounders X. We can distinguish two forms of departures from comparability:

- Incomplete overlap: the *support* of the distribution of X differs across the groups. I.e., some observations have no empirical counterfactuals. In these zones of no overlap, the model is forced to extrapolate, inferences are therefore based entirely on modeling assumptions instead of data.

[25] Independence removes selection bias *in expectation over all hypothetical trials*. Independence does not imply actual balance in any single trial (Deaton and Cartwright, 2018).

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

- Imbalance: the *shape* of the distribution of X differs across the groups. The simple difference of group averages then might not be a reliable estimate of the ATET.

Matching The least similar the distributions of X across the groups (the least overlap and balance), the more inferences rely on the model instead of data, thus the least robust they are to model misspecification. To alleviate this concern of needing to specify the form of the relationship correctly, we can *restructure* our sample prior to analysis, namely modify it such that it resembles one from a randomized trial. Indeed, if the two groups have sufficient overlap and balance, then, even if we misspecify the model, we should still get a reasonable estimate of the TE (Gelman et al., 2020). Matching procedures permit just that: we create matched groups to exhibit balance and overlap w.r.t. X.

⚠ *Matching provides more overlap and balance, not identification. It is not an alternative to a design-based method.*[27] *It is merely a first step that reduces the reliance on parametric assumptions of the regression model. For matching to be able to capture by itself a causal effect, all the difference between the groups would need to be captured by X. This assumption of "selection on observables" is very strong, and not testable. Therefore a matching estimator by itself, i.e., an identification strategy based solely on matching, is considered much less credible than one based on some exogenous variation. We need exogenous variation to believe the CIA.*[28] *With both, i.e., if (i) the CIA is satisfied, and (ii) we have balance and overlap w.r.t. X, then the difference in average observed outcomes is an unbiased estimate of the true ATET.*

Common matching methods

- Propensity score matching (PSM)
  Units are matched based on their predicted probability of getting treated, called their "propensity score". A logistic regression of D_i on X_i produces predicted probabilities of getting treated $\hat{p}_i$, which are then used to match each treated unit to non-treated units.

  ⚠ *PSM ensures $\hat{p}(X^t) = \hat{p}(X^c)$, but not $X^t = X^c$.*

  Algorithm for Estimating the Propensity Score (Doug's lecture notes) & Gelman et al. (2020) p. 420

- Mahalanobis Distance Matching
  Matching is based on a distance metric which can include multiple dimensions of "closeness" between observations: $\text{Distance}(X^c, X^t) = \sqrt{(X^c - X^t)' S^{-1} (X^c - X^t)}$ (whereas PSM reduces the dimensionality to 1). Each treated unit is matched to its nearest control unit, and control units are not reused.

  This method gets as close as possible to: $X^c = X^c \implies \hat{p}(X^T) = \hat{p}(X^C)$.

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 of the estimate, i.e., reduce its standard error.
- 👎 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

[27] 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).

[28] In other words, matching does not bring causality — nor do regression controls. Both are only adjustment strategies.

- 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_i with X.[29]

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.

"Falsification" or "placebo" tests are done almost automatically in RCTs — though rarely identified as such. In observational studies, the general approach is to estimate an alternative specification, which, under the identifying assumption, should not find an effect. An effect $\neq 0$ will suggest that the identifying assumption is violated, a confounder is probably driving the relationship.

Approach

- **RCT** The identifying assumption is random assignment (of each individual into the treatment vs control group). If treatment was indeed randomly assigned, then the sample means of explanatory variables should be the same across groups *in expectation*. RCT papers therefore typically show a "balance test" table of sample means of the Xs by group.[30][31]
- **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).

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

[30] ⚠ One shouldn't show t-tests (only balance tables) of baseline observables. t-tests in this context are conceptually unsound: they amount to assessing the probability of an event (a difference in averages) having occurred by chance, when we already know that it could only occur by chance, as the allocation between treatment arms was carried out randomly. 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."* t-tests are only sound in the sense of wanting to test empirically whether the randomization was carried out correctly.

[31] Doug Almond's advice: even in observational settings, *always* show a balance test table. Andrew Gelman's advice: to show (im)balance in averages of X across groups, plot the standardized and absolute differences in mean values for the continuous and binary Xs, respectively.

- **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_i = property value, D_i = 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 Mechanism & 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 study's 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...).*
   – Show a balance test table (table of sample means of the covariates X by treatment and control group), to check that the X are balanced across the treatment and control group, i.e. that X is uncorrelated with treatment.
4. Estimation method
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 stochastic nature and properties of the estimators arises from the stochastic nature of the assignment of the treatments, rather than a sampling-based perspective where the uncertainty arises from the random*[32] *sampling of units from a large population"* (Athey and Imbens, 2018). 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*,[33] 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:[34]

1. Re-assign treatment randomly, i.e., draw from the "randomization set"[35] (respecting the structure of the original assignment mechanism, e.g., within strata), thus generating fake treatment statuses.

[32] 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.

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

[34] 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.

[35] 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.[36] 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.[37]

[36] 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.

[37] 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.

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)

Read Ghanem and Smith (2021)

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

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

2.1.2 Expressing TE as a linear regression

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 $\mathbb{E}[\beta_i|D_i{=}1]$.

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

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

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

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

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

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

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

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

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

3.3 IV

IV estimand

$$\begin{aligned}\beta_{\text{IV}} \equiv \frac{cov[Y_i, Z_i]}{cov[D_i, Z_i]} &= \frac{\mathbb{E}[Y_i Z_i] - \mathbb{E}[Y_i]\mathbb{E}[Z_i]}{\mathbb{E}[D_i Z_i] - \mathbb{E}[D_i]\mathbb{E}[Z_i]} \\ &= \frac{\mathbb{E}[Y_i|Z_i{=}1]\mathrm{P}(Z_i{=}1) - \Big(\mathbb{E}[Y_i|Z_i{=}1]\mathrm{P}(Z_i{=}1) + \mathbb{E}[Y_i|Z_i{=}0]\mathrm{P}(Z_i{=}0)\Big)\mathrm{P}(Z_i{=}1)}{\mathbb{E}[D_i|Z_i{=}1]\mathrm{P}(Z_i{=}1) - \Big(\mathbb{E}[D_i|Z_i{=}1]\mathrm{P}(Z_i{=}1) + \mathbb{E}[D_i|Z_i{=}0]\mathrm{P}(Z_i{=}0)\Big)\mathrm{P}(Z_i{=}1)} \\ &= \frac{\mathbb{E}[Y_i|Z_i{=}1]\big(1 - \mathrm{P}(Z_i{=}1)\big) - \mathbb{E}[Y_i|Z_i{=}0]\mathrm{P}(Z_i{=}0)}{\mathbb{E}[D_i|Z_i{=}1]\big(1 - \mathrm{P}(Z_i{=}1)\big) - \mathbb{E}[D_i|Z_i{=}0]\mathrm{P}(Z_i{=}0)} \\ &= \frac{\mathbb{E}\big[Y_i|Z_i{=}1\big] - \mathbb{E}\big[Y_i|Z_i{=}0\big]}{\mathbb{E}\big[D_i|Z_i{=}1\big] - \mathbb{E}\big[D_i|Z_i{=}0\big]}\end{aligned}$$

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

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

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

- Denominator:

$$\begin{aligned}\mathbb{E}\big[D_i|Z_i{=}1\big] - \mathbb{E}\big[D_i|Z_i{=}0\big] &= \mathbb{E}\big[D_i^1 - D_i^0\big] \\ &= 1 \times P\big(D_i^1 - D_i^0{=}1\big) \;+\; 0 \times P\big(D_i^1 - D_i^0{=}0\big) \;- 1 \times P\big(D_i^1 - D_i^0{=}-1\big) \\ &= P\big(D_i^0{=}0, D_i^1{=}1\big) \text{ as the probability of defiers is 0}\end{aligned}$$

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

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:

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

3.4 RD

Sharp RD estimand

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

Fuzzy RD estimand

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

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

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

3.5 DiD, DiDD

DiD estimand

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

DiDiD estimand

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

References

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

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

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

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

Athey, S. and Imbens, G. W. Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. Working Paper, Sept. 2018. URL https://www.nber.org/papers/w24963.

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

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

Fisher, S. R. A. *The Design of Experiments*. Oliver and Boyd, 1935. URL https://play.google.com/store/books/details?id=-EsNAQAAIAAJ.

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

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

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

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

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

Gibson, J. Are You Estimating the Right Thing? An Editor Reflects. *Applied Economic Perspectives and Policy*, 41(3):329–350, 2019. doi: 10.1093/aepp/ppz009.

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

Kowalski, A. E. Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform. *The Review of Economics and Statistics*, pages 1–45, 07 2021. doi: 10.1162/rest_a_01069.

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

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

Neyman, J. 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, 1923. ISSN 1454-7376.

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

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

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

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