Causal Inference in Observational Studies

Theory, Methods and Presentation

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 effects of a specific event or situation ("treatment"), and sets internal validity as the priority[2], then one wants a research design that may credibly identify causal effects. That ideal research design is a **randomized trial**: an experiment which randomly assigns the participants to either a treatment or a control group.
This experiment can be considered as the "gold standard" against which to judge other research designs. Therefore, 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 X 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 controlable 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

2.1 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_{0i}, Y_{1i} the potential outcomes. The potential outcomes framework[3] allows us to define quantities:

–	Treatment effect (TE)	$Y_{1i} - Y_{0i}$	*what we want to estimate*
–	Average treatment effect (ATE)	$\mathbb{E}[Y_{1i} - Y_{0i}]$	*what we ideally want to compute*
–	Average treatment effect on the treated (ATET)	$\mathbb{E}[Y_{1i} - Y_{0i}\|D_i{=}1]$	*what we reasonably want to compute*
–	Difference in average observed outcomes	$\mathbb{E}[Y_i\|D_i{=}1] - \mathbb{E}[Y_i\|D_i{=}0]$	*what we can compute*
–	Difference in average observed outcomes for same X_i	$\mathbb{E}[Y_i\|D_i{=}1,X_i] - \mathbb{E}[Y_i\|D_i{=}0,X_i]$	*what we can compute*

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

- To measure TE = $Y_{1i} - Y_{0i}$, we need to see the same individual with and without treatment.
- This is impossible, as each individual has one existence. We cannot observe the counterfactual[4]. We can only compute the difference in average observed outcomes:

$$\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0] \;=\; \ldots \;=\; \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i{=}1]}_{\text{ATET}} \;+\; \underbrace{\mathbb{E}[Y_{0i}|D_i{=}1] - \mathbb{E}[Y_{0i}|D_i{=}0]}_{\text{selection bias}}$$

  The "selection bias" corresponds to the difference in Y_{0i} between the treated and untreated[5].
- If treatment is randomly assigned, then potential outcomes are independent of the treatment $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$, so there is no selection bias and ATET = ATE. The *independence assumption* identifies the TE.
- In observational studies, $(Y_{0i}, Y_{1i}) \not\perp\!\!\!\perp D_i$. However, if we *match* treated and control individuals to be proper counterfactuals, i.e., if the potential outcomes are *conditionally* independent of the treatment $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i|X_i$, then we again eliminate selection bias. The difference in average observed outcomes for treated and control individuals that share the same confounding covariate X_i equals the ATET: the *conditional independence assumption*[6] (CIA) identifies the TE.

> What makes our regression estimates causal is our **identifying assumption**:
>
> – if we have the IA $\quad (Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \quad \Longrightarrow$ we can compute the ATE.
>
> – if ~~IA~~ but we have the CIA $\quad (Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i|X_i \quad \Longrightarrow$ we can compute the ATET.
>
> – if ~~CIA~~ but $\exists$ a relevant instrument Z_i that is an exogenous source of variation in D_i:
>
> $$\begin{cases} \text{independent (new CIA)} & (Y_{0i}, Y_{1i}) \perp\!\!\!\perp Z_i|X_i \\ \text{relevant} & Z_i \not\perp\!\!\!\perp D_i|X_i \end{cases} \Longrightarrow \text{we can compute } a \text{ LATE.}$$
>
> 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) and was extended to observational studies by Rubin (1974). A strong assumption of the representation is that the treatment effect on one individual is independent of the treatment received by others. This excludes feedback effects and strategic interactions among agents.

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

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

[6] The CIA is also referred to as the assumption of "selection on observables" (conditioning on observed characteristics X_i erases selection bias), unconfoundedness, or "ignorability" in statistics.

Elaborate econometric techniques or regression controls won't bring causality. Reciprocally, with a good identification strategy, one need simply compare means — no *need* for controls.

2.2 Expressing TE as a linear regression

Suppose the TE is the same for everyone, i.e., $Y_{1i} - Y_{0i} = \beta$. The relation between observed outcomes and potential outcomes (how we estimate our TE) can be written as a linear regression on the treatment:

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

Consider the simple linear regression $Y_i = \alpha + \beta D_i + u_i$.

- 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.
- But $\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0] = ... = \beta + \mathbb{E}[u_i|D_i{=}1] - \mathbb{E}[u_i|D_i{=}0]$: it equals the true β only if the error u_i is uncorrelated with the treatment D_i.
- This correlation $\mathbb{E}[u_i|D_i{=}1] - \mathbb{E}[u_i|D_i{=}0]$ is equal to $\mathbb{E}[Y_{0i}|D_i{=}1] - \mathbb{E}[Y_{0i}|D_i{=}0]$: the difference in potential outcome Y_{0i} between the treated and the untreated, or "selection bias".

I.e., an identification problem (dependence) $\implies$ a regression problem (endogeneity).

2.3 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 the error term e, making e correlated with D.
- non-random (correlated to y) measurement error in D
- omitted variable bias (OVB): All omitted variables[7] 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 that we have measured all confounders[8], we turn to alternative causal inference or **"identification" strategies**, that rely on other assumptions.

[7] An omitted variable is an explanatory variable not included in the regression but which is a determinant of y.
[8] For instance, in cross-sectional approaches, we worry about time-invariant omitted variables. As a cross-section offers only

3 Applied Identification Methods

3.1 Hierarchy of common identification methods

A contestable hierarchy of the most common identification methods in the 'randomista' toolkit[9], based on their capacity to mimic random assignment, is as follows:

0. Randomized experiment (RCT)
1. Instrumental Variables (IV) and regression discontinuity (RD)
   If we fear that there is selection into treatment based on unobservables, we use an instrument that induces quasi-experimental variation in treatment status.
2. Difference-in-differences (DiD) and event-studies
   If we have repeated observations data and want to estimate the effect of an event, we use research designs that rely on the assumptions of parallel trends and time-invariant differences.
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 abaility to recover a causal effect – than strategies based on some exogenous variation. However, matching is a type of procedure that can complement a natural experiment design. It is adressed in section 4.

The sections below present, for each method, in the canonical setup, 1. the (assumed) data generating process (DGP), 2. the identifying assumptions, 3. the estimand of interest, 4. the estimator used, and 5. some best practices, and strengths and weaknesses.

Importantly, the **relation between the actual observed outcomes** Y_i (or Y_{it}) **and the conceptual potential outcomes** Y_{0i}, Y_{1i} 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.

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 fixed effects, we can focus on within variation only, which greatly reduces the threat of OVB.

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

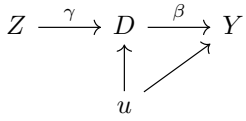
3.2 Random assignment

Show:

- experimental: RCT
- observational but still random assignment: Example with weather + another example

3.3 IV

Context (DGP) $Y_i = \alpha + \beta\ D_i + u_i, \quad cov[D_i, u_i] \neq 0$: D_i is endogenous. But $\exists$ a binary instrument Z_i that constitutes a source of variation in D_i, it "assigns treatment" or changes the probability of treatment[10]:



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

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_{0i}, which will be realized as 0 if individuals comply, 1 if not. $Z_i = 1$ induces D_{1i}, 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_{0i}	D_{1i}
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: $cov[Z_i, u_i] = 0$
} 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 homogenous effects.

Estimand We define the IV estimand: $\beta_{\text{IV}} \equiv \dfrac{cov[Y_i, Z_i]}{cov[D_i, Z_i]} = ... = \dfrac{\mathbb{E}[Y_i|Z_i{=}1] - \mathbb{E}[Y_i|Z_i{=}0]}{\mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0]}$. 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 $\widehat{\gamma.\beta}_{\text{LS}} = \frac{\widehat{cov}[Y_i, Z_i]}{\widehat{\mathbb{V}}[Z_i]}$ 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 $\dfrac{\mathbb{E}[Y_i|Z_i{=}1] - \mathbb{E}[Y_i|Z_i{=}0]}{\mathbb{E}[D_i|Z_i{=}1] - \mathbb{E}[D_i|Z_i{=}0]}$ to $\underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_{0i}{=}0, D_{1i}{=}1]}_{\textbf{LATE on the compliers}}$

Estimator Our natural choice of estimator is the sample analog $\widehat{\beta}_{IV} = \frac{\widehat{cov}[Y_i, Z_i]}{\widehat{cov}[D_i, Z_i]}$. It turns out to be numerically equivalent to the 2SLS estimator $\widehat{\beta}_{2SLS}$ obtained through the two-step procedure[11]:

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

[10] 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.

[11] 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.
- Count and characterize the compliers to get more out of the LATE (see section on external validity)

Strengths & weaknesses

+ Compelling identification strategy
– Strong assumptions
– The IV estimator will be less efficient than the OLS estimator if the instrument is weak.

3.4 RD (sharp)

Known assignment mechanism but no overlap

Context (DGP) Treatment D_i is not randomly assigned, it is assigned deterministically but *discontinuously* along Z_i, s.t. there is "local randomization" around a cutoff c. Units to the right of c receive treatment, those to the left don't ($D_i = \mathbb{1}\{Z_i \geq c\}$). The only confounder is the assignment variable Z_i.

$$Y_i = \alpha + \beta.D_i + k(Z_i) + u_i \ ^{12}$$

Identifying assumptions

(A1) *local* independence: $Y_{1i}, Y_{0i} \perp D_i | Z_i$ for $Z_i \in [c-a, c+a]$, i.e., no confounders vary discontinuously across the threshold. $\implies$ The average outcome of those right below the cutoff (who are denied the treatment) are a valid counterfactual for those right above (who receive the treatment).

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

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

Estimand $$\beta_{\text{RD}} = \lim_{z \to c^+} \mathbb{E}[Y_i | Z_i = z] - \lim_{z \to c^-} \mathbb{E}[Y_i | Z_i = z] \ = \ ... \ = \ \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | Z_i = c]}_{\textbf{LATE at the cutoff}}$$

Estimator We could estimate β by running two separate regressions on each side of c, after transforming Z to $Z-c$, s.t. the intercepts correspond to the values of the regression functions at c.
We can estimate β (and its SE) *directly* by running a pooled regression – including interaction terms between D and Z to let the regression function differ on both sides of the cutoff:

$$\left.\begin{aligned} Y_i &= \alpha_l + k_l(Z_i - c) + v_i \\ Y_i &= \alpha_r + k_r(Z_i - c) + w_i \end{aligned}\right\} \longrightarrow \begin{aligned} Y_i &= \alpha_l + (\alpha_r - \alpha_l)\mathbb{1}\{\text{right}\} + k(Z_i - c) + e_i \\ &= \alpha_l + \boldsymbol{\beta} D_i + k(Z_i - c) + e_i \end{aligned}$$

where $k() = k_l() + [k_r() - k_l()]D_i$

Best practices

- Choice of $k()$ and non-parametric estimation:
  - Option #1: local linear regression, with choice of the bandwidth h: $Y_i = \alpha_l + \boldsymbol{\beta} D_i + \gamma_l(Z - c) + (\gamma_r - \gamma_l)(Z - c)D + e_i$ with $c - h \leq X \leq c + h$. A larger h increases precision but also bias[13].
  - Option #2: local polynomial regression model with choice of the polynomial's order.

  In both cases, report the results of several specifications to assess the sensitivity to $k()$.

Strengths & weaknesses

+ RDDs are similar to a local randomized experiment, and thereby require weak assumptions.
+ 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.

[12] The function $k(.)$, and therefore the true functional form of the DGP, is unknown. This is a problem, as misspecification of the functional form typically biases the estimate. Estimation in RDDs is therefore generally done non-parametrically (using polynomial regression models, or local linear regression models).

[13] 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, the model is fit using only the observations $Z_i - h \leq \mathbf{Z} < Z_i < c$ when $Z_i < c$, and only the observations $c < Z_i < \mathbf{Z} \leq Z_i + h$ when $Z_i \geq c$.

3.5 DiD

Context (DGP) We want to estimate the causal effect of *an event*, which occurs at time t_0 and affects a group of individuals in the population, on some outcome Y.
Treatment assignment is a function of 2 dimensions: group (treatment/control) and period (pre/post event). We define the binary variables $G_i \equiv \mathbb{1}\{i \in \text{treatment group}\}$ and $P_t \equiv \mathbb{1}\{t \in \text{post period}\}$.

Identifying assumptions

(A1) Same counterfactual trends across groups, i.e., comparable outcomes in the absence of treatment:
$\mathbb{E}\big[Y_{0i}|G_i{=}1, P_t{=}1\big] - \mathbb{E}\big[Y_{0i}|G_i{=}1, P_t{=}0\big] = \mathbb{E}\big[Y_{0i}|G_i{=}0, P_t{=}1\big] - \mathbb{E}\big[Y_{0i}|G_i{=}0, P_t{=}0\big]$

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

Estimand

$$
\begin{aligned}
\beta_{\text{DiD}} &\equiv \big(\bar{Y}_{G_1P_1} - \bar{Y}_{G_1P_0}\big) - \big(\bar{Y}_{G_0P_1} - \bar{Y}_{G_0P_0}\big) \\
&\equiv \Big(\mathbb{E}\big[Y_{it} \mid G_i{=}1, P_t{=}1\big] - \mathbb{E}\big[Y_{it} \mid G_i{=}1, P_t{=}0\big]\Big) - \Big(\mathbb{E}\big[Y_{it} \mid G_i{=}0, P_t{=}1\big] - \mathbb{E}\big[Y_{it} \mid G_i{=}0, P_t{=}0\big]\Big) \\
&= \Big(\mathbb{E}\big[Y_{1i} \mid G_i{=}1, P_t{=}1\big] - \mathbb{E}\big[Y_{1i} \mid G_i{=}1, P_t{=}0\big]\Big) - \Big(\mathbb{E}\big[Y_{0i} \mid G_i{=}0, P_t{=}1\big] - \mathbb{E}\big[Y_{0i} \mid G_i{=}0, P_t{=}0\big]\Big) \\
&= \dots \\
&= \underbrace{\mathbb{E}\big[Y_{1i} - Y_{0i}|G_i{=}1, P_t{=}1\big]}_{\textbf{ATET}}
\end{aligned}
$$

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

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

If we have panel data (instead of merely repeated cross-sections), we can estimate this in a more direct way. As $\beta = \mathbb{E}[Y_{1i} - Y_{0i}|G_i = 1] - \mathbb{E}[Y_{1i} - Y_{0i}|G_i = 0]$, we can simply regress the change for each unit $Y_{1i} - Y_{0i}$ on G_i: a first difference approach.

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 effect of treatment actually *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.

Strengths & weaknesses

+ DiD is a way of ruling out unobserved time-invariant confounders, by using repeated observations, thus 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.

3.6 DiDiD

Context (DGP) The treatment varies along a 3rd dimension or "subgroup" (in addition to time and place), such as gender, season... We define the binary variable $S_i \equiv \mathbb{1}\{i \in \text{treatment subgroup}\}$.

Identifying assumptions

(A1) Same counterfactual trends across ~~groups~~ subgroups:

$$\begin{aligned}\big(\mathbb{E}[Y_{0i}|G_1,S_1,P_1] - \mathbb{E}[Y_{0i}|G_1,S_1,P_0]\big) - \big(\mathbb{E}[Y_{0i}|G_1,S_0,P_1] - \mathbb{E}[Y_{0i}|G_1,S_0,P_0]\big) \\ = \big(\mathbb{E}[Y_{0i}|G_1,S_1,P_1] - \mathbb{E}[Y_{0i}|G_1,S_1,P_0]\big) - \big(\mathbb{E}[Y_{0i}|G_1,S_0,P_1] - \mathbb{E}[Y_{0i}|G_1,S_0,P_0]\big)\end{aligned}$$

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

Estimand

$$\begin{aligned}\beta_{\text{DiDiD}} &\equiv \Big[\big(\bar{Y}_{G_1S_1P_1} - \bar{Y}_{G_1S_1P_0}\big) - \big(\bar{Y}_{G_0S_1P_1} - \bar{Y}_{G_0S_1P_0}\big)\Big] - \Big[\big(\bar{Y}_{G_1S_0P_1} - \bar{Y}_{G_1S_0P_0}\big) - \big(\bar{Y}_{G_0S_0P_1} - \bar{Y}_{G_0S_0P_0}\big)\Big] \\ &= \dots \\ &= \underbrace{\mathbb{E}\big[Y_{1i} - Y_{0i}|G_i{=}1, S_i{=}1, P_t{=}1\big]}_{\textbf{ATET}}\end{aligned}$$

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.

Strengths & weaknesses

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

3.7 Event study

Context (DGP) We want to estimate the causal effect of *an event*[14], which occurs at different times t_{0i} 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: the event is unpredictable, and not a result of the outcome Y.
Random timing of the event implies that there cannot be any pre-trends, s.t. we can reasonably use a unit's past value to construct its counterfactual post-event value.

Estimand $\beta_{\text{ES}} \equiv \mathbb{E}\big[Y_{it}|t{=}t_{0i}\big] - \mathbb{E}\big[Y_{it}|t{=}t_{0i}{-}1\big] = \ldots = \underbrace{\mathbb{E}\big[Y_{1i} - Y_{0i}|P_t{=}1\big]}_{\textbf{ATET}}$

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

$$Y_{it} = \sum_{\tau=-K}^{t_0-1} [\beta_\tau \mathbb{1}\{t=\tau\}] + \boldsymbol{\beta} \; \mathbb{1}\{t=t_0\} + \sum_{\tau=t_0+1}^{L} [\beta_\tau \mathbb{1}\{t=\tau\}] + \lambda_i + e_{it}$$

Best practices

- Report all β_τs, 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.

Strengths & weaknesses

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

[14] 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.

Summary of common identification methods

	Source of identification & identifying assumptions	**Estimand β & corresponding TE**	**Chosen estimator $\widehat{\beta}$**	**Strengths / Weaknesses**
RCT	(A) independence	$\beta_{\text{RCT}} \equiv \mathbb{E}[Y_i \mid D_i{=}1] - \mathbb{E}[Y_i \mid D_i{=}0] = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}]}_{\textbf{ATE}}$	$\widehat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \alpha + \boldsymbol{\beta}\ D_i + e_{it}$. Is **consistent** and **unbiased**.	+ Random assignment structurally guarantees (A) $\implies$ RCT = *"gold standard"*
IV	Id. from the exogenous variation in D induced by Z. (A1) independence (A2) exclusion restriction (A3) relevance (A4) monotonicity	$\beta_{\text{IV}} \equiv \dfrac{cov[Y_i, Z_i]}{cov[D_i, Z_i]} = \ldots$ $= \dfrac{\mathbb{E}[Y_i \mid Z_i{=}1] - \mathbb{E}[Y_i \mid Z_i{=}0]}{\mathbb{E}[D_i \mid Z_i{=}1] - \mathbb{E}[D_i \mid Z_i{=}0]}$: *"Wald estimand"* $= \ldots$ $= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i}{=}1, D_{0i}{=}0]}_{\textbf{LATE, compliers}}$	$\widehat{\beta}_{\text{IV}} \equiv \dfrac{\widehat{cov}[Y_i, Z_i]}{\widehat{cov}[D_i, Z_i]} = \ldots =$ numerically equivalent to $\widehat{\beta}_{\text{2SLS}}$ Is: **consistent, biased**, but bias $\downarrow$ 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 Z. (A1) independence (A2) relevance	$\beta_{\text{RD}} \equiv \lim_{z \to c^+} \mathbb{E}[Y_i \mid Z_i = z] - \lim_{z \to c^-} \mathbb{E}[Y_i \mid Z_i = z]$ $= \ldots$ $= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid Z_i = c]}_{\textbf{LATE, at the cutoff}}$	Estimate non-parametric regression functions: $Y_i = \alpha_l + \boldsymbol{\beta} D_i + k(Z_i - c) + e_i$ – local linear regression, bandwidth h – polynomial regression Is: **consistent, biased**, bias $\uparrow$ 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	$\beta_{\text{DiD}} \equiv \Big(\mathbb{E}[Y_{it} \mid G_i{=}1, P_t{=}1] - \mathbb{E}[Y_{it} \mid G_i{=}1, P_t{=}0]\Big)$ $- \Big(\mathbb{E}[Y_{it} \mid G_i{=}0, P_t{=}1] - \mathbb{E}[Y_{it} \mid G_i{=}0, P_t{=}0]\Big)$ $= \ldots$ $= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid G_i{=}1, P_t{=}1]}_{\textbf{ATET}}$	$\widehat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \alpha + \boldsymbol{\beta}\ G_i P_t + \lambda_G + \lambda_P + e_{it}$ Is **consistent**.	+ rules out unobserved time-invariant confounders
DiDiD	(A1) same counterfactual trends across subgroups (A2) same subgroup composition over time	$\beta_{\text{DiDiD}} \equiv \Big[\big(\bar{Y}_{G_1 S_1 P_1} - \bar{Y}_{G_1 S_1 P_0}\big) - \big(\bar{Y}_{G_0 S_1 P_1} - \bar{Y}_{G_0 S_1 P_0}\big)\Big]$ $- \Big[\big(\bar{Y}_{G_1 S_0 P_1} - \bar{Y}_{G_1 S_0 P_0}\big) - \big(\bar{Y}_{G_0 S_0 P_1} - \bar{Y}_{G_0 S_0 P_0}\big)\Big]$ $= \ldots$ $= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid G_i{=}1, S_i{=}1, P_t{=}1]}_{\textbf{ATET}}$	$\widehat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \alpha + \boldsymbol{\beta}\ G_i S_i P_t + \lambda_{GS} + \lambda_{GP} + \lambda_{PS} + e_{it}$ 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	$\beta_{\text{ES}} \equiv \mathbb{E}[Y_{it} \mid t{=}t_0] - \mathbb{E}[Y_{it} \mid t{=}t_0{-}1]$ $= \ldots$ $= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid P_t{=}1]}_{\textbf{ATET}}$	$\widehat{\beta}_{\text{OLS}}$ of the regression $Y_{it} = \boldsymbol{\beta}\ \mathbb{1}\{t = t_{0i}\} + \sum_{\tau \neq \{t_{0i}-1, t_{0i}\}} [\beta_\tau \mathbb{1}\{t = \tau\}] + \lambda_i + e_{it}$ Is **consistent**.	– difficult to rule out unobserved confounders

4 Improving our causal inference

4.1 *Pre-estimation = improving design* – Matching

Matching refers to procedures that restructure the original sample prior to statistical analysis. The goal of this restructuring in a causal inference setting is to create a sample analogous to one created from a randomized experiment, i.e., we want the matched groups to be balanced and overlap[15] w.r.t. the pre-treatment variables that we consider to be confounders X. This may involve 1:1 matching, weighting, or subclassification.

What does matching provide? Recall that if the independence assumption is satisfied conditional on the *confounding*[16] covariates: $(Y_0, Y_1) \perp\!\!\!\perp D|X$, and if we can achieve balance and overlap w.r.t. them, then the observed difference in means across treatment and control groups $\mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_1|D = 0]$ will be an *unbiased* estimate of the true treatment effect (TE).

Matching should not be thought of as an alternative to a *research design*-based method[17], but as a first step in the causal analysis that can reduce the reliance on parametric assumptions of the model fitted to estimate the TE. Just as with a randomized experiment, the intuition is that if the treatment and control groups have sufficient overlap and balance, then, even if we misspecify the model, we should still get a reasonable estimate of the treatment effect (Gelman et al., 2020).

For matching to have the ability to capture by itself a *causal* effect, the assumption of selection on observables would have to be satisfied, i.e., that all the difference between the groups is captured by X. This is a very (too) strong assumption, that is 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. Matching does not bring causality; nor do regression controls. Both are only adjustment strategies.

Common matching methods

- Propensity score matching (PSM)
  Units are matched based on their "propensity score": their predicted probability of getting treated.
  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 $X^t = X^c$.

  Methods that stratify the data should be preferred.

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

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

[15] The treatment and control groups can be different in two ways:
- Imbalance: the distributions of the confounders differ across the groups. → The simple difference of group averages is not a good estimate of the ATE. We must adjust for pre-treatment differences, e.g., by matching or weighting.
- Incomplete overlap: the support (range of x) differs across the two groups. We have no empirical counterfactuals for some treatment/control observations. The model will create counterfactual predictions by extrapolating over portions of the space where there is no data to support them.

[16] The pre-treatment variables that predict (= are associated with) both treatment assignment and the outcome.

[17] 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).

4.2 *Estimation* – Good/bad regression controls

Separately from the identification strategy, controlling for balance covariates that are unrelated to D_i (and influence y) can increase the efficiency of the estimate (reduce the residual variation); whereas covariates that are correlated with D_i will introduce bias.

In addition to adding the X as predictors, Gelman et al. (2020) recommends considering adding interactions of the treatment D_i with those X that have large estimated effects, to look at how the treatment effect varies with the level of X.

Finally, if we center X, the treatment coefficient represents the treatment effect for individuals with the mean X score for the sample.

Penalizing complexity in linear regressions and variable selection:

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

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

  It overcomes the limitations of LASSO. If $\lambda = 0$, this is OLS; if $\alpha = 0$, this is LASSO.
  Suresh 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.

4.3 *Post-estimation = checking assumptions* – Falsification Tests

In empirical studies concerned with causal inference, one can never directly *test* the **identification assumptions**, but one can do falsification analyzes to *support* their validity – and thus the **internal validity** of the study.

They do not *prove* that the assumptions hold, instead their results either reject the claim, or increase confidence in it.

They are done almost automatically in RCTs – though rarely identified as such. They are virtually as easy to do in observational studies; the general approach is to estimate an alternative specification, which, if the identifying assumption holds, should not find an effect. Finding an effect $\neq 0$ will suggest the identifying assumption is violated.

Approach

- **RCT** The identifying assumption is random assignment (of each individual into the treatment vs control group). If that's true, then the sample means of explanatory variables should be the same across groups. A "balance test" table of sample means of the Xs by group is therefore a falsification test of an RCT's central assumption[18][19].
- **IV** The two main identifying assumptions can be tested:
  - relevance (Z_i is strongly related to sorting into treatment D_i): directly observable in the 1st stage;

[18] ⚠ One should show balance tables (not t-tests!) 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.

[19] 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 differences in mean values for the Xs.

- exclusion restriction (Z_i isn't correlated with Y_i through some pathway other than D_i). The ideal falsification test is to estimate the reduced form effect of Z_i on Y_i in a situation where Z_i can't affect D_i. Finding an effect means Z_i affects Y_i through an other channel than D_i, 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_i and Y_i).

- **RD** The two main identifying assumptions can be tested:
  - continuity or "local randomization" (all other factors determining Y_i evolve "smoothly" w.r.t. Z_i). 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_i on Z_i: $D_i = \mathbb{1}\{Z_i \geq c\}$). 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 *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 are based on 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? Would the results hold?*

In causal inference studies, there is also another type of variation at play: variation in *assignment of treatment*. There is "design-based" uncertainty induced from not knowing 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[20] *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? Would the results hold?*

RI is a method for calculating p-values for hypothesis tests. RI may also be used for construction of confidence intervals, but this application requires stronger assumptions.

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 H0 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*[21], we can simply use $\hat{\beta}$ as the test-statistic and obtain its empirical distribution under H0. Indeed:

- If there is no effect for any unit, then a unit's potential outcomes are the same: the observed outcome is also the counterfactual. Under H0, 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 a reference distribution under H0.
- 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.

Sampling-based inference	Randomization inference
H0	
No average effect: $\mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] = 0$	"Sharp" no effect: $Y_{1i} - Y_{0i} = 0, \forall i$
Distribution of $\hat{T}$ under H0	
Under H0, the distribution of $\hat{T}$ across all random samples converges (as $n \to \infty$) to a known distribution: usually t or normal. $\longrightarrow$ We compute the parameters of this distribution. $\longrightarrow$ The asymptotic distribution of $\hat{T}$ (across random samples) = the "sampling distribution".	Under H0, however 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 many simulated assignments. $\longrightarrow$ The distribution of $\hat{T}$ (across random treatment assignments) = the "randomization distribution".
***2-sided p-value* = Pr [observing a $\hat{T} > \lvert\hat{T}_{\text{obs}}\rvert$] under H0** **= share of the distribution that is $> \lvert\hat{T}_{\text{obs}}\rvert$**	
= Pr[*the observed difference between groups would have been observed*] if they had been drawn from underlying sampling frames with no mean difference.	= Pr[*the observed difference between groups would have been observed*] if they had received a different treatment assignment.
$\Longrightarrow$ Given e.g. a rejection threshold $\alpha = 0.05$, the test will erroneously reject H0 < 5% of the time	

In practice: simulation We repeat a large number of times (e.g., 10000) the following procedure[22]:

1. Re-assign treatment randomly, i.e., draw from the "randomization set"[23] (respecting the structure of the original assignment mechanism, e.g., within strata), thus generating fake treatment statuses.

[20] 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.

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

[22] 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.

[23] 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.

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[24].
- RI allows us to make inferences about causal effects without having to appeal to large samples, i.e., without relying on asymptotics, as it is based on the act of physical randomization alone.
  - When all possible random assignments can be simulated, the reference distribution is known, thus RI produces exact p-values.
  - When the set of possible random assignments is so large that a complete census of possible random assignments is infeasible, the reference distribution can be approximated by randomly sampling from the randomization set many times.
- 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[25].

[24] 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.

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

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 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?

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...

Counting and characterizing compliers to get more out of a LATE Compliers are rarely representative, due to selective uptake. While we cannot identify individual compliers in the data, we can estimate the size of the complier group: $P[D_{i1} > D_{i0}] = \mathbb{E}[D_{i1} - D_{i0}] = \mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$, and characterize them in terms of their distribution of observed covariates (see Kowalski (2018), Abadie (2003)).

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.

6 Other branches of causal modelling

6.1 Structural Equation Models (SEMs)

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

SEMs[26] 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 modelling approaches:

1. Implicitly assumes that the relationships among variables (paths) are causal.
   A big leap from the traditional statistics' "correlation does not imply causation". By using pre-existing knowledge of the system, one can make an informed hypothesis about the causal structure of the variables, and SEM explicitly tests this supposed causal structure[27].
2. Variables can be both predictors and responses. → SEM is useful for testing and quantifying indirect (cascading) effects – that would otherwise go unrecognized by any single model.

Traditional SEM

Estimation: Coeffs 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 (data has no underlying structure)
- Normal errors (model fit to normal distrib)

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 v-cov 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. Piece the m paths together after for inferences about the entire SEM. (we estimate m separate vcov matrices.)

$\Longrightarrow$ 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: $AIC = C + 2k\frac{n}{n-k-1}$

6.2 Graphical causal modeling

[26] SEMs are a fast growing statistical technique in ecological research. A new way to study ecological systems.
[27] Causality is central: SEM is designed to test competing hypotheses about complex relationships. On causality in SEMs: Judea Pearl (2012) and Kenneth Bollen & Judea Pearl (2013).

Keep reading: https://mixtape.scunning.com/dag.html#a-simple-dag

A Maths of potential outcomes

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

2.1 The original selection bias problem

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

2.2 Expressing TE as a linear regression

Consider the simple linear regression $Y_i = \alpha + \beta D_i + u_i$.

- The OLS slope estimand simplifies to the difference in average observed outcomes:

$$\begin{aligned}\beta_{\text{OLS}} &= \frac{cov[Y_i, D_i]}{Var[D_i]} = \frac{\mathbb{E}[Y_i D_i] - \mathbb{E}[Y_i]\mathbb{E}[D_i]}{\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2}\\ &= \frac{\mathbb{E}[Y_i{\times}1|D_i{=}1]\mathrm{P}(D_i{=}1) - \Big(\mathbb{E}[Y_i|D_i{=}0]\mathrm{P}(D_i{=}0) + \mathbb{E}[Y_i|D_i{=}1]\mathrm{P}(D_i{=}1)\Big)\Big(\frac{1}{2}{\times}1 + \frac{1}{2}{\times}0\Big)}{\Big(\frac{1}{2}\times 1^2 + \frac{1}{2}\times 0^2\Big) - \Big(\frac{1}{2}\times 1 + \frac{1}{2}\times 0\Big)^2}\\ &= \frac{\mathbb{E}[Y_i|D_i{=}1]\times\frac{1}{2} - \Big(\mathbb{E}[Y_i|D_i{=}0]\times\frac{1}{2} + \mathbb{E}[Y_i|D_i{=}1]\times\frac{1}{2}\Big)\times\frac{1}{2}}{\frac{1}{2} - \frac{1}{4}}\\ &= \frac{\mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0]\times\frac{1}{2} - \mathbb{E}[Y_i|D_i{=}1]\times\frac{1}{2}}{\frac{1}{2}}\\ &= \mathbb{E}[Y_i|D_i{=}1] - \mathbb{E}[Y_i|D_i{=}0]\end{aligned}$$

- However this is equal to the true β only if the error term is uncorrelated with the treatment:

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

3.3 IV

The IV estimand is defined as:

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

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

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

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

- Denominator:

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

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

References

Abadie, A. Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics, 113(2):231–263, 2003. ISSN 0304-4076. doi: 10.1016/S0304-4076(02)00201-4.

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.

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

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.

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. Working Paper 24647, National Bureau of Economic Research, May 2018. URL http://www.nber.org/papers/w24647.

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.

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.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol., 66(5):688–701, Oct. 1974. ISSN 0022-0663, 1939-2176. doi: 10.1037/h0037350.