13 Causal Inference

A substantive portion of empirical work in economics is interested in finding a causal effect of one variable – denoted in the following by d_i – on another variable y_i for a particular population of units (e.g. individuals). This is typically done in the context of the impact evaluation of a policy d_i on an outcome y_i, such as the impact of a new drug on health outcomes, the impact of a worker-retraining programme on labor market outcomes, or the impact of some subsidy on firms' sales. Thereby, d_i is called the treatment variable, y_i the outcome variable, and we refer to a causal effect also as a treatment effect and to the process of finding it as the identification problem. We can define a causal effect of d_i on y_i as the change in y_i that would materialize if we changed d_i and held everything else constant.

Causal questions are ceteris-paribus-questions: they ask what would happen with the outcome y_i if we changed the treatment d_i and kept everything else equal. As such, they are also what-if-questions: they consider a hypothetical, counterfactual state of the world that did not materialize. Because of that, individual treatment effects can (typically) not be identified, and even average treatment effects for a specific subpopulation can never be identified from the data alone. Rather, the researcher needs to supply identifying assumptions and convince the audience of their credibility.

The previous chapters dealt with structural (or parametric) econometric models, which specify a parametric functional form for $\mathbb{E}[y_i|d_i]$ and therefore the potential outcomes of unit i under different treatments. The causal effect implied by structural models correctly identifies a causal effect only if the specified model is correct, which is typically a very strong assumption and renders the analysis non-credible. Rather than to evaluate the impact of a particular real-world policy, structural analysis can be used to estimate a causal effect within the context of an economic theory, provided that the particular functional form of the model is tightly derived from an economic theory.[1]

[1] However, structural models are typically used for other goals than estimating a causal effect. For

This chapter discusses the identification of causal effects with experimental and quasi-experimental empirical methods. They explicitly state the causal effect of interest in terms of counterfactual outcomes and aim at identifying an average causal effect for some population using methods that mimic an ideal, often infeasible experiment that would randomly assign different treatments to different units. The chapter begins in Section 13.1 with a detailed exposition of the potential outcomes framework, a definition of the causal effects of interest and a comparison to the linear regression, a structural model. In turn, it discusses how causal effects can be identified using experiments in Section 13.2 and using quasi-experimental methods in Section 13.3. Importantly, the chapter leaves out another important approach to uncovering causal effects, which, rather than relying on potential outcomes and quasi-experimental methods and aiming for point-identification, uses minimal, credible assumptions to set-identify a causal effect of interest (see e.g. Manski (1995, 2008)).

13.1 Potential Outcomes Framework

The potential outcomes framework and experimental and quasi-experimental approaches to causal inference date back to Rubin (1974) and Robins (1986). They consider a unit i (e.g. individual), and define a random variable (RV) d_i *at the level of unit* i with possible realizations $d_i \in \{0,1\}$, indicating whether unit i received some treatment ($d_i = 1$) or not ($d_i = 0$). We seek to identify the causal effect of treatment d_i on some outcome y_i. For this purpose, we define y_i as a RV with possible realizations $y_i \in \{y_{0i}, y_{1i}\}$. This notation emphasizes that y_{di} is the realization of the outcome y_i that would materialize if unit i received treatment $d_i = d$, i.e.

$$y_i = \begin{cases} y_{0i} & \text{if } i \text{ received treatment, i.e. } d_i = 0 \\ y_{1i} & \text{if } i \text{ did not receive treatment, i.e. } d_i = 1 \end{cases} .$$

The difference in potential outcomes $y_{1i} - y_{0i}$ is referred to as the individual treatment effect. These potential outcomes are constants *at the level of unit* i. They are defined in a way that, *at the level of unit* i, the stochasticity of y_i only arises due to the stochasticity of d_i. In other words, provided that we know the possible realizations y_{0i} and y_{1i}, the premise is that, given the realization d of the treatment-indicator d_i, we can perfectly tell the realization of the

example, one might be interested in the partial correlation of y_i and a variable in x_i, holding fixed the other variables in x_i, in informing a structural parameter featured in an economic theory or in predicting y_i based on a set of available covariates x_i for observations i not contained in the sample. Thereby, besides economic theory, the model choice can also be justified by mathematical results. For example, in the context of time series data, the Wold decomposition leads to a General Linear Process, which can be approximated by an Autoregressive (AR) process of finite order (see Section 9.2).

outcome y_i, namely y_{di}. As a result, we can write

$$y_i = d_i y_{1i} + (1 - d_i) y_{0i} \ .$$

In principle, by virtue of being (discrete) RVs, both d_i and y_i each have a probability function, which, together with their possible realizations, defines various moments. However, their unconditional probabilities and moments *at the level of unit* i are not of interest. Only the conditional probability function for y_i given $d_i = d$ is of interest: it is a pointmass at y_{di}.[2]

This is a detailed statement of the Stable Unit Treatment Value Assumption (SUTVA). Essentially, SUTVA ensures that the individual treatment effect $y_{1i} - y_{0i}$ is a well-defined constant *at the level of unit* i and that it can be interpreted as the causal effect of changing d_i from 0 to 1. Environments that are incompatible with the one presented so far constitute violations of SUTVA. This occurs, for example, if for some units we observe y_{0i} even under treament (e.g. a patient avoids taking some administered drug without our knowledge), or if the potential outcomes (y_{0i}, y_{1i}) are a function of the treatment of other units. Reasons for the latter can be contagion (e.g. the vaccination of some other individual j ($d_j = 1$) improves the health outcome of individual i in absence of their own vaccination, y_{0i}), displacement (due to increased police in some cities j ($d_j = 1$), crime in other cities i, y_{0i}, goes up) or communication (workers i who did not participate in a retraining programme learn from the workers j who did ($d_j = 1$), affecting their y_{0i}).[3]

Definition of Causal Effects The RVs d_i and y_i with their two possible realizations, respectively, are defined for many units i in a supposedly infinite population. Whereas the possible realizations of d_i are the same for all observations, the possible realizations of y_i – the potential outcomes of i, (y_{0i}, y_{1i}) – are specific to each unit i. As a result, when looking *across units* i, not only d_i and y_i, but also y_{0i} and y_{1i} are all RVs. Each of them has a probability function, which, together with their possible realizations, defines various moments. In particular, there are different possible realizations (values) of y_{0i} and y_{1i} across i, which give rise to many possible realizations (values) of y_i across i. The expectations $\mathbb{E}[y_{0i}]$

[2] In principle, we could write this as follows: $y_i|(d_i = d) = y_{di}$ for $d \in \{0, 1\}$, i.e. the RV y_i, when conditioning on the realization $d_i = d$ of the RV d_i, is equal to y_{di}. However, presentations of the potential outcomes framework avoid using this notation with an explicit conditioning set and conditioning-line "|" because it is reserved for a different purpose, as explained further below. Nevertheless, it is important to emphasize that, because y_{di} is a constant *at the level of unit* i, it does not make sense to condition it on the realization of d_i: we would get $y_{0i}|(d_i = d) = y_{0i}$ regardless of the realization d of d_i, and similarly $y_{1i}|(d_i = d) = y_{1i}$.

[3] In such cases, one can often redefine the unit of analysis so as to satisfy SUTVA. For example, if one suspects such spillover effects among individuals only within the same household, one can do the analysis at the household- rather than individual-level. Another remedy is to compare units which are distant enough so as to exclude contagion, displacement and communication between units.

and $\mathbb{E}[y_{1i}]$ denote the average of these potential outcomes *across units* i in the population.[4]

Because $d_i \in \{0, 1\}$, each unit in the infinite population has either $d_i = 0$ or $d_i = 1$, and there are infinitely many treated ($d_i = 1$) units as well as infinitely many non-treated ($d_i = 0$) units in the population. We know that for the treated subpopulation, $y_i = y_{1i}$ holds, whereas for the non-treated subpopulation, $y_i = y_{0i}$ holds. However, in principle, the RV y_{0i} is defined for all units, including the ones in the treated subpopulation, and likewise y_{1i} is defined even for units in the non-treated subpopulation. As a result, while the expectations

$$\mathbb{E}[y_i] \; , \quad \mathbb{E}[y_{0i}] \; , \quad \text{and} \quad \mathbb{E}[y_{1i}]$$

denote, respectively, the averages of the outcome y_i, the potential outcome without treatment y_{0i} and the potential outcome with treatment y_{1i} across all units i in the population, the conditional expectations

$$\mathbb{E}[y_i|d_i = 1] \; , \quad \mathbb{E}[y_{0i}|d_i = 1] \; , \quad \text{and} \quad \mathbb{E}[y_{1i}|d_i = 1] = \mathbb{E}[y_i|d_i = 1]$$

denote the averages of the outcome y_i, the potential outcome without treatment y_{0i} and the potential outcome with treatment y_{1i} only for the treated subpopulation, and analogously the conditional expectations

$$\mathbb{E}[y_i|d_i = 0] \; , \quad \mathbb{E}[y_{0i}|d_i = 0] = \mathbb{E}[y_i|d_i = 0] \; , \quad \text{and} \quad \mathbb{E}[y_{1i}|d_i = 0]$$

do so for the non-treated subpopulation. Depending on how treatment is assigned in the overall population, these two subpopulations can be quite distinct from one another. In particular, they can differ with regard to the distributions of potential outcomes. This means that $\mathbb{E}[y_{0i}|d_i = 1]$ and $\mathbb{E}[y_{0i}|d_i = 0]$ and, consequently, $\mathbb{E}[y_{0i}]$ are in general not the same,[5] and likewise for expectations of y_{1i}. This allows us to define the Average Treatment Effect (ATE), the Average Treatment Effect for the Treatment-Group (ATT) and the Average Treatment Effect for the Control-Group (ATC) as distinct objects:

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[y_{1i} - y_{0i}] \; , \\
\text{ATT} &= \mathbb{E}[y_{1i} - y_{0i}|d_i = 1] \; , \\
\text{ATC} &= \mathbb{E}[y_{1i} - y_{0i}|d_i = 0] \; .
\end{aligned}$$

[4] Under violations of SUTVA, y_{1i} could be a RV *at the level of unit* i, which means that we could compute its expectation *at the level of unit* i. Dealing with such cases would require us to be explicit about the distribution w.r.t. which some expectation is computed, potentially defining several expectation operators. However, under SUTVA, y_{0i} and y_{1i} are constants *at the level of unit* i, and all expectation operators used in context of the potential outcomes framework denote averages *across units* i.

[5] Note that $\mathbb{E}[y_{0i}] = \mathbb{P}[d_i = 0]\mathbb{E}[y_{0i}|d_i = 0] + \mathbb{P}[d_i = 1]\mathbb{E}[y_{0i}|d_i = 1]$.

The ATE is the average of the individual treatment effect $y_{1i} - y_{0i}$ across all units i in the population, ATT is the average within the treated subpopulation ("treatment group"), and ATC is the average within the non-treated subpopulation ("control group"). Typically, we evaluate the effects of a policy that actually took place, which means that we are most interested in the ATT, i.e. we seek to know what difference the treatment made for the units who actually were treated.

Identification of Causal Effects Now suppose we observe treatments and outcomes for a random sample of n units from the overall population, $\{d_i, y_i\}_{i=1}^n$. Each unit i has either $d_i = 0$ or $d_i = 1$ and, correspondingly, either $y_i = y_{0i}$ or $y_i = y_{1i}$. As a result, we in fact observe $\{d_i, y_{d_i i}\}_{i=1}^n$. Let $\mathcal{N}_0 = \{i : d_i = 0\}$ and $\mathcal{N}_1 = \{i : d_i = 1\}$ be the sets of units in our sample who received and did not receive treatment, respectively, with sizes $n_0 = |\mathcal{N}_0|$ and $n_1 = |\mathcal{N}_1|$ such that $n = n_0 + n_1$. The above means that, while we observe a sample of size n of realizations of d_i and y_i from the overall population of all units, we observe a sample of size n_0 of realizations of y_{0i} from the non-treated subpopulation and a sample of size n_1 of realizations of y_{1i} from the treated subpopulation. Based on this data, we can use the analogy principle to consistently estimate the first term in the ATT formula and the second term in the ATC formula:

$$\begin{aligned}
\frac{1}{n_1}\sum_{i\in\mathcal{N}_1} y_i = \frac{1}{n_1}\sum_{i\in\mathcal{N}_1} y_{1i} \quad &\xrightarrow{p} \quad \mathbb{E}[y_{1i}|d_i = 1] = \mathbb{E}[y_i|d_i = 1] \ , \\
\frac{1}{n_0}\sum_{i\in\mathcal{N}_0} y_i = \frac{1}{n_0}\sum_{i\in\mathcal{N}_0} y_{0i} \quad &\xrightarrow{p} \quad \mathbb{E}[y_{0i}|d_i = 0] = \mathbb{E}[y_i|d_i = 0] \ .
\end{aligned}$$

Without further assumptions, we cannot identify the remaining terms. First, we cannot identify $\mathbb{E}[y_{0i}|d_i = 1]$ and $\mathbb{E}[y_{1i}|d_i = 0]$ because we do not observe y_{0i} for treated units ($i \in \mathcal{N}_1$) and we do not observe y_{1i} for non-treated individuals ($i \in \mathcal{N}_0$). Second, we cannot identify $\mathbb{E}[y_{1i}]$ and $\mathbb{E}[y_{0i}]$ because $\mathcal{N}_1$ is a random subset of the treated subpopulation, but not of the overall population, and likewise for $\mathcal{N}_0$.[6] As a result, the ATE is in general not identified from our data![7]

Ideally, we could observe $\{d_i, y_{0i}, y_{1i}\}_{i=1}^n$, i.e. both potential outcomes for each unit i, along with treatment d_i. In this case we could estimate all six objects that appear in ATE, ATT

[6] This is why the above two objects do not converge to $\mathbb{E}[y_{1i}]$ and $\mathbb{E}[y_{0i}]$, respectively.

[7] It would be tempting to conclude, based on the expression $y_i = d_i y_{1i} + (1 - d_i)y_{0i}$, that $y_{di} = y_i|(d_i = d)$ and in turn to argue that $\mathbb{E}[y_{1i}] = \mathbb{E}[y_i|d_i = 1]$ and to claim that the ATE is identified. Here it is important to be precise about the meaning of the conditioning-line "|". The potential outcomes framework uses it to distinguish different subpopulations, i.e. to distinguish quantities *across units* i, rather than to denote, *at the level of unit* i, the conditioning of one RV, like y_i, on the realization of another, like d_i.

and ATC using their respective sample analogues.[8] And not only that; such data would immediately give us individual treatment effects $y_{1i} - y_{0i}$. However, this is at odds with reality. The fact that we never observe both potential outcomes for the same unit is referred to as "the fundamental problem of causal inference". In absence of observing both potential outcomes, we can at most identify an average of the individual treatment effect $y_{1i} - y_{0i}$ for a particular population – such as the treated (ATT), non-treated (ATC) or even more granular subpopulations (see Sections 13.2 and 13.3) –, and even this only under identification assumptions. We aim at doing so in a way that allows us to interpret the resulting quantity as a causal effect: if we were to gather many units from that particular population and administer treatment $d_i = 1$ to them, we would expect (on average) to see a change in y_i equal to that quantity.

Under an ideal experiment, treatment d_i is assigned independently of potential outcomes (y_{0i}, y_{1i}). For example, say an experimental drug is randomly assigned to some individuals, while others receive a placebo. Then

$$\mathbb{E}[y_{1i}|d_i = 1] = \mathbb{E}[y_{1i}|d_i = 0] = \mathbb{E}[y_{1i}] \quad \text{and} \quad \mathbb{E}[y_{0i}|d_i = 1] = \mathbb{E}[y_{0i}|d_i = 0] = \mathbb{E}[y_{0i}] \; .$$

In turn, ATE, ATT and ATC coincide and can be identified from the data using a simple difference in average observed outcomes between the treatment- and control groups in our sample:

$$\frac{1}{n_1}\sum_{i \in n_1} y_i - \frac{1}{n_0}\sum_{i \in n_0} y_i \;\; \xrightarrow{p} \;\; \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] = \text{ATE} = \text{ATT} = \text{ATC} \; . \tag{13.1}$$

This environment and attempts of creating it are discussed in Section 13.2.

Typically, however, we deal with observational – rather than experimental data –, and we cannot exclude that units in the treated subpopulation ("treatment group") have significantly different potential outcomes than units in the non-treated subpopulation ("control group"), i.e. that d_i and (y_{0i}, y_{1i}) are correlated. In this more general case, the naive difference from

[8] Concretely,

$$\frac{1}{n_1}\sum_{i \in n_1} y_{0i} \;\xrightarrow{p}\; \mathbb{E}[y_{0i}|d_i = 1] \; , \quad \frac{1}{n_0}\sum_{i \in n_0} y_{0i} \;\xrightarrow{p}\; \mathbb{E}[y_{0i}|d_i = 0] \; , \quad \frac{1}{n}\sum_{i=1}^{n} y_{0i} \;\xrightarrow{p}\; \mathbb{E}[y_{0i}] \; ,$$
$$\frac{1}{n_1}\sum_{i \in n_1} y_{1i} \;\xrightarrow{p}\; \mathbb{E}[y_{1i}|d_i = 1] \; , \quad \frac{1}{n_0}\sum_{i \in n_0} y_{1i} \;\xrightarrow{p}\; \mathbb{E}[y_{1i}|d_i = 0] \; , \quad \frac{1}{n}\sum_{i=1}^{n} y_{1i} \;\xrightarrow{p}\; \mathbb{E}[y_{1i}] \; .$$

above can be written as

$$\begin{aligned}\mathbb{E}[y_i|d_i=1]-\mathbb{E}[y_i|d_i=0] &= \mathbb{E}[y_{1i}|d_i=1]-\mathbb{E}[y_{0i}|d_i=0]\\ &= \mathbb{E}[y_{1i}|d_i=1]-\mathbb{E}[y_{0i}|d_i=1]+\mathbb{E}[y_{0i}|d_i=1]-\mathbb{E}[y_{0i}|d_i=0]\\ &= ATT+\mathbb{E}[y_{0i}|d_i=1]-\mathbb{E}[y_{0i}|d_i=0]\ ,\end{aligned}$$

i.e. we obtain the ATT plus a term that reflects a selection bias; it shows the difference in the expectations of the potential outcome in absence of treatment, y_{0i}, between the treatment- and control-groups. For example, workers who end up participating in a retraining program ($d_i = 1$) might be the ones who would earn less in absence of participation (i.e. have lower y_{0i}) than the ones who did not participate in the program. In that case, the bias above would be negative, i.e. the naive difference would underestimate the ATT because it compares the earnings of workers who participated in the program to the earnings of workers who did not participate, and the latter are higher on average than the participating workers' hypothetical earnings that would have been obtained had they not participated.[9] With observational data, we aim at finding a way to avoid such a bias and identify the average treatment effect for some subpopulation using methods that *mimic* an ideal experiment. Such methods are discussed in Section 13.3.

Comparison to Linear Regression Drawing parallels between non-parametric causal analysis methods and parametric econometric models is not trivial because the former start from potential outcomes (y_{0i}, y_{1i}), use them to define the observed outcome y_i by relying on SUTVA and define an average treatment effect as the object of interest, whereas parametric models directly start with a model for the observed outcome y_i, without defining potential outcomes. Nevertheless, the selection bias above can be viewed as the non-parametric counterpart to regressor endogeneity in linear regression models (see Section 3.5). Consider the regression with an intercept and d_i as the only covariate:

$$y_i = \beta_0 + \beta_1 d_i + u_i\ , \tag{13.2}$$

[9] We can write the naive difference also as

$$\begin{aligned}\mathbb{E}[y_i|d_i=1]-\mathbb{E}[y_i|d_i=0] &= \mathbb{E}[y_{1i}|d_i=1]-\mathbb{E}[y_{0i}|d_i=0]\\ &= \mathbb{E}[y_{1i}|d_i=1]-\mathbb{E}[y_{1i}|d_i=0]+\mathbb{E}[y_{1i}|d_i=0]-\mathbb{E}[y_{0i}|d_i=0]\\ &= ATC+\mathbb{E}[y_{1i}|d_i=1]-\mathbb{E}[y_{1i}|d_i=0]\ ,\end{aligned}$$

i.e. the ATC plus a term that reflects the selection bias in terms of potential outcomes under treatment, y_{1i}. In principle, we only require independence of d_i and y_{0i} to estimate ATT, and we only require independence of d_i and y_{1i} to estimate ATC.

where $\mathbb{E}[u_i] = 0$ is w.l.o.g. because an intercept is included. The simple comparison of mean observed outcomes for the treatment- and control-groups yields

$$\begin{aligned}\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0] &= \mathbb{E}[\beta_0 + \beta_1 d_i + u_i|d_i = 1] - \mathbb{E}[\beta_0 + \beta_1 d_i + u_i|d_i = 0] \\ &= \beta_1 + \mathbb{E}[u_i|d_i = 1] - \mathbb{E}[u_i|d_i = 0] \; , \qquad (13.3)\end{aligned}$$

i.e. we obtain β_1, the parameter that we would be tempted to interpret as a causal effect of d_i on y_i,[10] as well as a selection bias-term $\mathbb{E}[u_i|d_i = 1] - \mathbb{E}[u_i|d_i = 0]$.[11] As discussed in Section 3.5, regressor endogeneity can be remedied by including the omitted variables in u_i into the regression (using the correct functional form) or by using instrumental variables (IVs). The experimental and quasi-experimental methods discussed in the following are motivated precisely by the fact that it is often impossible to correctly control for all omitted variables, while it can be hard to find good IVs. Nevertheless, the two approaches to remedy regressor endogeneity permeate these methods.

13.2 Finding Causality with Experimental Data

An experiment or Randomized Controlled Trial (RCT) randomly assigns treatment to units.[12] In the ideal case, we have full compliance, i.e. each unit i assigned to the treatment group actually receives treatment ($d_i = 1$) and each unit assigned to the control group does not receive treatment ($d_i = 0$). As explained in the previous section and stated in Eq. (13.1), this renders treatment d_i independent of any other RV – including in particular the potential outcomes (y_{0i}, y_{1i}) – which in turn renders ATE, ATT and ATC identified by the naive difference

$$\begin{aligned}\widehat{\text{ATE}} = \widehat{\text{ATT}} = \widehat{\text{ATC}} &= \frac{1}{n_1}\sum_{i \in n_1} y_i - \frac{1}{n_0}\sum_{i \in n_0} y_i \\ &\xrightarrow{p} \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] = \text{ATE} = \text{ATT} = \text{ATC} \; .\end{aligned}$$

To verify compliance with the randomly assigned treatment, one can check whether various background characteristics of units (e.g. age, income, sex, etc. in the case of individuals) are

[10] One might – rightfully – ask what the object of interest is in this environment. Typically, when practitioners write down linear regressions in the context of impact evaluation, they seek conditions under which β_1 can be interpreted as a causal effect.

[11] Note that treatment effect heterogeneity is not a reason for regressor endogeneity. If the true model, allowing for individual treatment effects, is $y_i = \beta_0 + \alpha_i d_i + e_i$, we can define $\beta_1 = \mathbb{E}[\alpha_i|d_i = 1]$ and write instead $y_i = \beta_0 + \beta_1 d_i + u_i$ with $u_i = b_i d_i + e_i$ and $b_i = \alpha_i - \beta_1$. Because $\mathbb{E}[u_i|d_i = 0] = \mathbb{E}[e_i|d_i = 0]$ and $\mathbb{E}[u_i|d_i = 1] = \mathbb{E}[b_i + e_i|d_i = 1] = \mathbb{E}[e_i|d_i = 1]$, the selection bias in the naive difference above depends on the correlation of d_i with e_i, the error term left after accounting for treatment effect heterogeneity.

[12] See Bertrand and Mullainathan (2004) or Fehr and Goette (2007) for examples of RCTs in economics.

balanced between the treatment and control groups, as they should be under randomized treatment.[13]

The above estimator is non-parametric, as it was derived without specifying a structural model that features parameters, like linear regressions feature β and σ^2. However, it turns out to be equivalent to the parametric OLS estimator $\hat{\beta}_1$ for the parameter β_1 in Eq. (13.2).[14] This is not necessarily true in other settings than the present one with randomized, binary treatment.[15]

Sometimes, other variables are added as separate covariates to that regression in Eq. (13.2) to reduce the variance of the estimator $\hat{\beta}_1$. This can be justified because such covariates do not change the probability limit of $\hat{\beta}_1$, as any covariate is uncorrelated with the randomized d_i.[16] The variance reduction occurs because such covariates explain part of the outcome variable, thereby reducing the variation left to be explained by d_i, whereas they cannot explain any variation in the random d_i. See Section 3.6 for details.

Imperfect Compliance The ideal experiment, in which units perfectly comply with our randomized treatment-assignment, is rare. Instead, under partial non-compliance, we explicitly distinguish treatment-assignment – denoted by $z_i \in \{0,1\}$ – and actual treatment or treatment-status – denoted by $d_i \in \{0,1\}$. We observe $\{y_i, d_i, z_i\}_{i=1}^n$. There are four different types of units in the population (see Table 13.1): always-takers obtain treatment regardless of whether they are assigned to the treatment- or control-group, never-takers avoid treatment regardless of assignment, compliers comply with our assignment ($d_i = z_i$), and defiers do the opposite of what is intended ($d_i = 1 - z_i$).[17] Typically, it is assumed that there are

[13] In the context of impact evaluation, variables other than y_i and d_i are referred to as units' "background characteristics".

[14] As derived in the Appendix to Chapter 3, we have $\hat{\beta}_1 = \frac{\sum_{i=1}^n d_i(y_i - \bar{y})}{\sum_{i=1}^n d_i(d_i - \bar{d})}$. Under binary treatment, $\sum_{i=1}^n d_i = \sum_{i=1}^n d_i^2 = n_1$. Furhter, define $\bar{y}_1 \equiv \frac{1}{n_1}\sum_{i=1}^n d_i y_i$ and $\bar{y}_0 \equiv \frac{1}{n_0}\sum_{i=1}^n (1-d_i) y_i = \frac{n}{n_0}\bar{y} - \frac{n_1}{n_0}\bar{y}_1$. We then get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n d_i(y_i - \bar{y})}{\sum_{i=1}^n d_i(d_i - \bar{d})} = \frac{n_1\bar{y}_1 - n_1\bar{y}}{n_1 - n_1^2/n} = \frac{\bar{y}_1 - \bar{y}}{n_0/n} = \bar{y}_1 - \left(\frac{n}{n_0}\bar{y} - \frac{n_1}{n_0}\bar{y}_1\right) = \bar{y}_1 - \bar{y}_0 \ .$$

[15] Very much related is the fact that the structural assumption $y_i = \beta_0 + \beta_1 d_i + u_i$ with $\mathbb{E}[u_i|d_i] = 0$ – i.e. $\mathbb{E}[y_i|d_i] = \beta_0 + \beta_1 d_i$ – is without loss of generality when d_i is binary and randomized. As a result, the ATE, ATT and ATC coincide with the parameter β_1, whereby $\mathbb{E}[u_i|d_i] = \mathbb{E}[u_i] = 0$ holds.

[16] See the derivation of the probability limit of $\hat{\beta}$ under omitted variables in Section 3.5.2.

[17] Constructing the counterfactual treatment status and distinguishing four types of units is helpful to motivate the identified treatment effect below. More formally, one could define d_{0i} and d_{1i} as the treatment status that would materialize for unit i under assignment to the control- ($z_i = 0$) or treatment-group ($z_i = 1$), respectively. Just as only one of the two potential outcomes (y_{0i}, y_{1i}) is observed, so too we observe only a single treatment status d_i out of (d_{0i}, d_{1i}) for each unit because each unit is either assigned to the treatment- or the control-group.

no defiers, which is referred to as the monotonicity assumption.[18]

Table 13.1: Experiment Compliance

	treatment-status when assigned to	
	treatment-group ($z_i = 1$)	control-group ($z_i = 0$)
Always-Takers	yes ($d_i = 1$)	yes ($d_i = 1$)
Never-Takers	no ($d_i = 0$)	no ($d_i = 0$)
Compliers	yes ($d_i = 1$)	no ($d_i = 0$)
Defiers	no ($d_i = 0$)	yes ($d_i = 1$)

The table shows the (hypothetical) treatment status d_i as a function of treatment assignment z_i.

In this environment, the naive difference

$$\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0]$$

does not necessarily identify ATE, ATT and ATC because, while treatment-assignment z_i is random, treatment-status d_i might not be. This is because the two terms above not only include compliers, who were randomized into receiving ($d_i = 1$) or not receiving treatment ($d_i = 0$), but the first term also includes always-takers, who received treatment despite our assignment to the control-group, and the second term includes never-takers, who avoided treatment despite our assignment to the treatment-group. Individuals in those two groups are likely different with regard to their potential outcomes than compliers, rendering the treated and non-treated populations different and causing a selection bias in the naive difference. For example, workers who opt out of a retraining programme to which they were assigned might on average have better outside options y_{0i}.

The intention-to-treat effect (ITT) is the difference in average outcomes between the treatment- and control-groups:

$$\text{ITT} = \mathbb{E}[y_i|z_i = 1] - \mathbb{E}[y_i|z_i = 0] ,$$

which is identified. Under full compliance, all units are compliers (i.e. $d_i = z_i$ for all i) and the ITT coincides with the naive difference

$$\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0] ,$$

which under randomized treatment coincides with ATE, ATT and ATC.

Under partial non-compliance, the ITT does not identify ATE, ATT and ATC because of the presence of always- and never-takers (presuming the monotonicity assumption holds),

[18] In other words, the effect of z_i on d_i is assumed to be non-negative for everyone.

which drives the ITT towards zero. More concretely, because of never-takers, the first term in the ITT contains non-treated units, and because of always-takers, the second term in the ITT contains treated units.[19] [20] We can correct for this downward bias by scaling up the ITT by the fraction of compliers in our sample, $\mathbb{E}[d_i|z_i = 1] - \mathbb{E}[d_i|z_i = 0] = \mathbb{P}[d_i = 1|z_i = 1] - \mathbb{P}[d_i = 1|z_i = 0]$.[21] In this way we get the average treatment effect for the compliers, i.e. for the units for whom being assigned to the treatment- as opposed to the control-group actually pushes them into obtaining treatment:

$$\text{LATE} = \frac{\mathbb{E}[y_i|z_i = 1] - \mathbb{E}[y_i|z_i = 0]}{\mathbb{E}[d_i|z_i = 1] - \mathbb{E}[d_i|z_i = 0]} \ .$$

It is a particular type of a Local Average Treatment Effect (LATE), whereby "local" refers to compliers in this case. A numerical example is instructive. Suppose the average outcomes for the treatment- and control-groups are 15 and 5, respectively. Suppose further that only 25% of units in the treatment-group actually received treatment, whereas no one in the control-group received treatment (one-sided non-compliance, without always-takers). The ITT is 10. However, intuitively, this is only a fourth of the actual treatment effect because a change in z_i from 0 to 1 induces on average an increase in d_i of only 0.25. To get the effect of a full increase of d_i from 0 to 1, we multiply the ITT by 4 (or divide it by 0.25) to arrive at an estimated LATE of 40. A plausible example of such one-sided non-compliance without always-takers is the assignment of an experimental drug that cannot be obtained unless one is assigned to the treatment-group. An example of two-sided non-compliance is when z_i indicates whether individual i obtained a voucher to pay for tuition at some university or private high school, whereby many individuals enrol even without vouchers and typically not all individuals with vouchers enrol.

We can estimate the LATE non-parametrically by replacing the expectations with sample means:

$$\widehat{\text{LATE}} = \left(\frac{1}{n_1^z} \sum_{i \in \mathcal{N}_1^z} y_i - \frac{1}{n_0^z} \sum_{i \in \mathcal{N}_0^z} y_i \right) \Big/ \left(\frac{1}{n_1^z} \sum_{i \in \mathcal{N}_1^z} d_i - \frac{1}{n_0^z} \sum_{i \in \mathcal{N}_0^z} d_i \right) \ ,$$

where $\mathcal{N}_1^z$ and $\mathcal{N}_0^z$ denote the sets of individuals assigned to the treatment- and control groups, respectively, with sizes n_1^z and n_0^z. Once again, in this simple setting with binary treatment and binary, randomized treatment assignment, one can show that it coincides with

[19] In the limit, as the fraction of compliers in the population decreases to zero, the ITT becomes zero, because the randomized treatment- and control-groups contain the same fractions of always- and never-takers, respectively.

[20] Because z_i is randomized, the respective fractions of always- and never-takers should be balanced among the treatment- and control-groups.

[21] The former term includes always-takers and compliers, while the latter includes always-takers and, in principle, defiers, of which there are none under the monotonicity assumption.

the parametric two-stages least squares (2SLS) Instrument Variable (IV) estimator, in which treatment-assignment z_i is used as an IV for treatment-status d_i. [22] With this in mind, we can add covariates to the 2SLS regression to obtain a more efficient estimator of LATE.

This LATE is the average treatment effect for the subpopulation of compliers. Because the latter's distribution of potential outcomes can differ from that of the overall population, the LATE differs from the ATE. For example, students for whom a tuition voucher is critical for enroling in a private high school are a particular subpopulation of all students, and we expect the average effect enrolment in a particular high school on grades or other outcomes to be different for this subpopulation than for all students taken together

Further Considerations The discussion above points to the more general question of internal vs external validity. Even if a quantity of interest is identified, say in an ideal experiment, it typically teaches us something about the effect of that policy intervention only for settings that resemble the one of our experiment. With partial non-compliance in particular, the LATE might be different when looking at different samples of students to whom we randomly assign vouchers, and, depending on how effective we were in identifying and assigning vouchers to (potential) compliers, we can get very different effects of voucher assignment on outcomes (ITT).[23] Because of these considerations, it is important to design experiments that replicate well the treatment intervention setting of interest.

A potential problem in estimating causal effects with experimental data is attrition, i.e. missing outcome data for some units. For example, some people might drop out of an experimental drug study or some people might not fill out a final survey that measures outcomes. Attrition is rarely random, and if it is indeed systematically related to potential outcomes, then ignoring it yields biased estimates. One potential remedy is to assume that attrition is random once we control for other covariates x_i. If so, we can compute Conditional Average Treatment Effects (CATE) (see Section 13.3.3) and thereby ignore attrition.[24] Another possible approach under attrition is to compute bounds for the identified treatment effect. Suppose the outcome variable is bounded between 0 and 100. Filling in a zero for missings from the treatment group and 100 for missings from the control group and computing the treatment effect on this synthetic data yields a lower bound for the actual treatment effect, and doing the reverse yields an upper bound. For unbounded outcome

[22] Again, very much related is the fact that the structural assumptions $\mathbb{E}[y_i|d_i] = \beta_0 + \beta_1 d_i$ and $\mathbb{E}[d_i|z_i] = \gamma_0 + \gamma_1 z_i$ are without loss of generality when d_i and z_i are binary. As shown in the Appendix, the LATE then coincides with β_1 from the second-stage regression $y_i = \beta_0 + \beta_1 \hat{d_i} + e_i$, where $\hat{d_i} = \alpha_0 + \alpha_1 z_i$.

[23] Such considerations are particularly important if treatment assignment is costly.

[24] We can test whether attrition is driven by the characteristics x_i we observe by regressing a dummy for missing outcome variables on these characteristics and looking at the R^2. However, we cannot test whether attrition is driven by outcome variables.

variables, one could use the lowest and highest observed values of the outcome variable to construct the synthetic data.

13.3 Finding Causality with Observational Data

Oftentimes, conducting a field experiment by randomly assigning units to treatment- and control-groups is infeasible. For example, it is impossible to randomize exposure to dictatorial regimes in childhood to measure its impact on various attitudes in adulthood, and due to ethical concerns we cannot randomize neighborly noisyness to study its impact on, say, stress levels.[25] In those cases, we are left with observational rather than experimental data on outcomes y_i and treatments d_i, whereby d_i is endogenous. There are two related and interacting approaches to remedy this endogeneity: (i) find a natural experiment, i.e. an exogenous event or policy that causes as-good-as-random variation in treatment-status d_i for some subjects, (ii) find treated and non-treated units that are similar enough (e.g. by relying on other variables x_i) to argue that treatment-status among these units is as-good-as-random.

13.3.1 Instrumental Variables

In Section 13.2, the randomized treatment-assignment z_i is used as an IV to isolate the exogenous variation in the endogenous treatment-status d_i and find the latter's effect on outcomes, avoiding the endogeneity issues that would arise when measuring directly the effects of d_i on y_i. This very same framework can also be used for natural experiments. In that context, we search for an IV z_i that affects the treatment of interest d_i and, though not random, is plausibly independent of other factors that affect the outcome variable y_i.[26] An example where the IV is indeed random, though not randomized by the researcher in the context of an RCT, is Angrist (1990)'s study of how having served in the military (veteran status) impacts subsequent lifetime earnings. To isolate the exogenous variation in veteran status, he uses the lottery by which the US government in 1970 randomly assigned some young men to serve in the Vietnam war.

Another famous and illustrative example of a study using an IV to identify causal effects in the context of natural experiments is Angrist and Krueger (1991). They use individuals'

[25] Nevertheless, in those cases it is still useful to formulate what the ideal experiment would be. This highlights the (potential) endogeneity issues and paves the path for finding a method to solve them.

[26] In other words, we search for an IV that satisfies the relevance and exogeneity conditioons (see Section 7.4). In the context of impact evaluation, this setting is referred to as a natural experiment to emphasize the link to the ideal case of an actual experiment. In the latter, the experimental IV z_i refers to treatment-assignment and is randomized. The notion of a natural experiment shall highlight that the variation in the observational IV is as-good-as-random.

quarter of birth as an IV for educational attainment to measure its impact on subsequent earnings. They exploit the fact that, in the US, individuals' age at which they start school is affected by their birthdate, whereas the age at which going to school is no longer compulsory is set to 16, creating a variation in the compulsory schooling time between individuals born in different times of the year. At the same time, they argue that individuals' birthdate has no direct impact on earnings.

This study illustrates well the fact that, in the context of IV-estimation, the LATE measures the average treatment effect only for compliers: the IV of Angrist and Krueger (1991) allows them to find the effect of schooling on earnings only for the individuals whose educational attainment is actually influenced by the variation in the minimum years of schooling due to birthdates, i.e. individuals who drop out of school as soon as they are legally allowed to. This surely is a very particular subpopulation, meaning that the LATE tells us little about the ATE in the population. While in RCTs, the randomized treatment-assignment is the only IV available to the researcher and gives rise to a single complier-population,[27] under natural experiments there are in principle many different IVs that a researcher can take to measure the causal effect of a given treatment on a given outcome. Each of these IVs leads to a particular complier-population, for whom average treatment effects are computed. Choosing between different IVs – and estimation approaches more generally – may involve a trade-off between credibility and generality, as the most credibly exogenous IVs may involve a rather peculiar complier population.

A further contrast to the use of IVs in RCTs is that under natural experiments the IV z_i is rarely random, but instrument exogeneity has to be argued-for. To defend it, one might need to condition the analysis on covariates x_i, leading to Conditional LATEs (see Section 13.3.3).

A final difference of the use of IVs in natural experiments to their use in RCTs is that IVs in natural experiments oftentimes have a low power, i.e. explain a small portion of the variation in d_i. In RCTs, this is typically not the case, as there is a substantive complier-population. In contrast, under natural experiments, the as-good-as-random IV that nature provides us with might explain only a very small part of the variation in treatments, just as is the case in Angrist and Krueger (1991)'s use of quarter-of-birth as an instrument for educational attainment. Weak IVs yield estimators whose final sample distribution is far from Normal, hence calling into question the usual procedures for hypothesis testing and confidence interval construction (see Section 7.4).

[27] The researcher has an immediate influence on this complier population through the experimental design.

13.3.2 Regression Discontinuity Design

Sometimes, treatment d_i is determined based on whether some variable q_i crosses some threshold c: $d_i = \mathbf{1}\{q_i \geq c\}$. An example would be the eligibility for reduced public transport fares after a certain age, the passing of an entrance exam with more than a certain threshold of points, or the election of a candidate if they receive more than 50% of votes. The Regression Discontinuity Design (RDD) exploits the fact that such thresholds are often arbitrary and argues that treated units who just passed the threshold are comparable to non-treated units who just missed the threshold in the sense that treatment among those units is as-good-as-random. As a result, comparing average outcomes y_i of treated units with $q_i \in \mathcal{Q}_+ \equiv [c, c+\epsilon]$ and non-treated units with $q_i \in \mathcal{Q}_- \equiv [c-\epsilon, c]$ identifies a LATE among units with $q_i \in \mathcal{Q} \equiv \mathcal{Q}_+ \cup \mathcal{Q}_- = [c-\epsilon, c+\epsilon]$:

$$\begin{aligned}\text{LATE} &= \mathbb{E}[y_i|q_i \in \mathcal{Q}_+] - \mathbb{E}[y_i|q_i \in \mathcal{Q}_-] \\ &= \mathbb{E}[y_i|d_i = 1, q_i \in \mathcal{Q}] - \mathbb{E}[y_i|d_i = 0, q_i \in \mathcal{Q}] \ .\end{aligned}$$

For example, to determine how the usage of cars is impacted by making public transport free, we could exploit a rule by which a city offers free fare for, say, individuals older than 75 years. Comparing the car usage of individuals aged 75-76 to that of individuals aged 74-75, we could argue to have uncovered the average effect of free public transport on car usage for 74-76 year-olds.[28]

We can estimate this LATE by

$$\widehat{\text{LATE}} = \frac{1}{n_{1,\mathcal{Q}}} \sum_{i \in \mathcal{N}_1 \cap \mathcal{N}_\mathcal{Q}} y_i - \frac{1}{n_{0,\mathcal{Q}}} \sum_{i \in \mathcal{N}_0 \cap \mathcal{N}_\mathcal{Q}} y_i \ ,$$

where $\mathcal{N}_1$ and $\mathcal{N}_0$ denote, as before, the sets of treated and non-treated individuals, respectively, $\mathcal{N}_\mathcal{Q} = \{i : q_i \in \mathcal{Q}\}$ denotes the set of individuals that are close enough to the threshold. As a result, $\mathcal{N}_1 \cap \mathcal{N}_\mathcal{Q}$ and $\mathcal{N}_0 \cap \mathcal{N}_\mathcal{Q}$ denote the sets of treated and non-treated individuals, respectively, around the threshold, with sizes $n_{1,\mathcal{Q}} = |\mathcal{N}_1 \cap \mathcal{N}_\mathcal{Q}|$ and $n_{0,\mathcal{Q}} = |\mathcal{N}_0 \cap \mathcal{N}_\mathcal{Q}|$.

Note that this LATE is local in a different sense than the IV-based LATE from Sections 13.2 and 13.3.1, namely because it computes the average treatment effect for units near the threshold. Also, under this binary d_i, $\widehat{\text{LATE}}$ coincides with the OLS estimator of β_1 from the regression

$$y_i = \beta_0 + \beta_1 d_i + u_i \quad \text{for} \ \ i \in \mathcal{N}_\mathcal{Q} \ ,$$

while LATE coincides with the parameter β_1.

[28] See Angrist and Lavy (1999) or Eugster et al. (2017) for examples of studies applying RDDs in economics.

For an RDD to be plausible, we require the distribution of potential outcomes to be the same for units on either side of the threshold (local exclusion restriction).[29] In the above example, this would be violated if the free fare is offered as soon as individuals reach their retirement age, as retirement likely influences car usage by other channels than just reduced public transport fares. Even in cases where the threshold is indeed arbitrary, the local exclusion restriction becomes more credible as we zoom-in more towards the threshold, i.e. as we decrease ϵ. For example, an RDD comparing the car usage of individuals who are one year away from the threshold of age 75 gives a more convincing LATE than one doing so for individuals 5 years away from the threshold, because car usage, the general desire for mobility and possibly even attitudes towards public transport differ between individuals of different age and belonging to different cohorts. However, the effect for 70-80 year olds is surely more relevant than the effect for 74-76 year olds. This implies once again a trade-off between credibility and generality. It also implies a trade-off between credibility and estimation precision, as zooming-in leaves ever less observations for estimation. Using covariates, we can ameliorate this trade-off, i.e. render the exclusion restriction more credible and increase the estimation precision. This includes the forcing-variable q_i, which oftentimes directly affects outcomes y_i, but also other covariates x_i. In the example from before, it is a good idea to control for individuals' age, but also other measures of their desire and ability to be mobile. Thereby, it is important to assess the sensitivity of results to different functional form assumptions for q_i in order to avoid mistaking a nonlinear but smoothly changing effect of q_i on y_i around the threshold for a discontinuity that is interpreted as a treatment effect.

"Fuzzy" RDD The above discussion and example refers to a sharp RDD, where a forcing-variable q_i crossing a threshold c unambiguously determines treatment d_i. This is plausible in some cases, as with a sharp cutoff of 50% of votes required to be elected or free public transport for anyone older than 75. In other cases, when the forcing-variable crosses the threshold, this might lead to a discontinous increase in the treatment probability rather than causing the latter to jump from 0 to 1.

In such cases, we can conduct a fuzzy RDD, which involves estimating the LATE with $z_i = \mathbf{1}\{q_i \geq c\}$ as an instrument for d_i:

$$\text{LATE} = \frac{\mathbb{E}[y_i|q_i \in \mathcal{Q}_+] - \mathbb{E}[y_i|q_i \in \mathcal{Q}_-]}{\mathbb{E}[d_i|q_i \in \mathcal{Q}_+] - \mathbb{E}[d_i|q_i \in \mathcal{Q}_-]} = \frac{\mathbb{E}[y_i|z_i = 1, q_i \in \mathcal{Q}] - \mathbb{E}[y_i|z_i = 0, q_i \in \mathcal{Q}]}{\mathbb{E}[d_i|z_i = 1, q_i \in \mathcal{Q}] - \mathbb{E}[d_i|z_i = 0, q_i \in \mathcal{Q}]} \,.$$

The assumption thereby is that z_i is (as-good-as) random for individuals near the threshold, $i \in \mathcal{N}_{\mathcal{Q}}$. The non-parametric esitmator of the above LATE is constructed analogously to the

[29] One can check whether units around the threshold are similar in terms of background characteristics x_i, but one cannot check whether they are similar in terms of the distributions of potential outcomes.

$\widehat{\text{LATE}}$ in Section 13.2:

$$\widehat{\text{LATE}} = \left(\frac{1}{n^z_{1,\mathcal{Q}}} \sum_{i \in \mathcal{N}^z_1 \cap \mathcal{N}_\mathcal{Q}} y_i - \frac{1}{n^z_{0,\mathcal{Q}}} \sum_{i \in \mathcal{N}^z_0 \cap \mathcal{N}_\mathcal{Q}} y_i \right) / \left(\frac{1}{n^z_{1,\mathcal{Q}}} \sum_{i \in \mathcal{N}^z_1 \cap \mathcal{N}_\mathcal{Q}} d_i - \frac{1}{n^z_{0,\mathcal{Q}}} \sum_{i \in \mathcal{N}^z_0 \cap \mathcal{N}_\mathcal{Q}} d_i \right) .$$

All comments from Section 13.3.1 also apply to fuzzy RDDs.[30] In addition, this LATE is in two ways local: it computes the treatment effect only for those units at the threshold whose treatment-status is indeed affected by q_i crossing the threshold. For example, suppose individuals aged over 75 are not offered free public transportation, but only a reduced fare, and suppose they are not automatically eligible for the reduction, but have to claim a reduction authorization that is monthly renewable. In that case, conducting a fuzzy RDD allows us to estimate the effect of reducing public transport fares on car usage only for individuals aged around 75 who actually engage in the effort to claim their monthly reduction authorization.

13.3.3 Conditional Treatment Effects & Matching Methods

As discussed in previous sections, under random treatment, there is no need to make use of covariates in the analysis from a consistency and credibility point of view, but one might control for such covariates in the regression of outcomes on treatment in order to obtain a more precise (parametric) estimator of average treatment effects. The same holds for IV-based approaches when the IV is random or exogenous. This section discusses how covariates can be used for two additional reasons: first, to account for heterogeneity by computing conditional treatment effects, and second, to increase the credibility of the analysis by softening the exogeneity assumption and requiring exogeneity of treatment or IVs only conditional on covariates.

Before discussing common approaches to make use of covariates in nonparametric or semi-parametric causal inference approaches, it is useful to elaborate on how covariates are used in the parametric linear regression model. Recall from Section 13.2 that, under random treatment d_i, the ATE, ATT, ATC are equal to $\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0]$ and coincide with the parameter β_1 from the regression

$$y_i = \beta_0 + \beta_1 d_i + u_i \ . \tag{13.4}$$

Similarly, from Section 13.3.2 we know that if treatment is random for units near some

[30] As in Section 13.2, provided that our instrument $z_i = \mathbf{1}\{q_i \geq c\}$ and treatment d_i are binary, we can obtain $\widehat{\text{LATE}}$ also parameterically by using 2SLS estimation but focusing only on individuals near the threshold, $i \in \mathcal{N}_\mathcal{Q}$.

threshold, i.e. for i with $q_i \in \mathcal{Q}$, the LATE $\mathbb{E}[y_i|d_i = 1, q_i \in \mathcal{Q}] - \mathbb{E}[y_i|d_i = 0, q_i \in \mathcal{Q}]$ coincides with β_1 from a regression as above, defined for i with $q_i \in \mathcal{Q}$. From a regression point of view, one could include covariates x_i in these regressions, i.e. estimate

$$y_i = \beta_0 + \beta_1 d_i + x_i' b + e_i \ , \tag{13.5}$$

to increase the precision of β_1. Furthermore, it is often argued that doing so ameliorates a selection bias that arises due to endogeneity of d_i. However, in order for the estimator for β_1 from Eq. (13.5) to be unbiased while the one from Eq. (13.4) is biased, we need (i) x_i to have a linear effect on y_i (so that $e_i = u_i - x_i' b$ is independent of x_i), and (ii) d_i to be independent of e_i (i.e. all other factors that affect y_i besides d_i and x_i), but not of $u_i = x_i' b + e_i$, which happens if d_i and x_i are not independent.[31] It is important not to forget the first condition, which is quite restrictive; only if x_i affects y_i linearly does a simple inclusion of x_i in the regression eliminate the selection/omitted variable bias. In addition to including covariates as separate regressors, applications of linear regressions often use them to construct and include interaction terms like $d_i \ \mathbf{1}\{x_i \in \mathcal{X}_*\}$:

$$y_i = \beta_0 + \beta_1 d_i + \alpha d_i \ \mathbf{1}\{x_i \in \mathcal{X}_*\} + v_i \ . \tag{13.6}$$

This yields different effects of d_i on y_i for units with $x_i \in \mathcal{X}_*$ than for units with $x_i \notin \mathcal{X}_*$, i.e. it yields conditional treatment effects, e.g. for young vs old. Adding more such interaction effects, $\sum_{s=1}^{S} \alpha_s d_i \ \mathbf{1}\{x_i \in \mathcal{X}_s\}$, one can get more nuanced heterogeneity of treatment effects. Using analogous arguments to before, one can argue that the estimator for β_1 from Eq. (13.6) is unbiased while the one from Eq. (13.4) is not if (i) v_i is independent of x_i (i.e. one included all the interaction terms necessary to model the true extent of the heterogeneity in treatment effects) and (ii) d_i is independent of v_i, i.e. of all factors that affect y_i other than d_i and x_i.

In causal inference methods, the use of covariates mirrors the inclusion of interaction terms rather than the "controlling for covariates" in linear regressions. Under the Conditional Independence Assumption (CIA), treatment d_i is independent of potential outcomes (y_{1i}, y_{0i}) conditional on covariates x_i. As a result, we have

$$f_1(x_i) \equiv \mathbb{E}[y_{1i}|x_i] = \mathbb{E}[y_{1i}|x_i, d_i = 1] = \mathbb{E}[y_{1i}|x_i, d_i = 0]$$

and

$$f_0(x_i) \equiv \mathbb{E}[y_{0i}|x_i] = \mathbb{E}[y_{0i}|x_i, d_i = 1] = \mathbb{E}[y_{0i}|x_i, d_i = 0] \ .$$

[31] Replacing independence by uncorrelatedness, the analogous statement can be made about consistency instead of unbiasedness.

The resulting Conditional Average Treatment Effect (CATE) is identified because both $y_{1i}|(x_i, d_i = 1)$ and $y_{0i}|(x_i, d_i = 0)$ are observed:

$$\text{CATE}(x_i) \equiv f_1(x_i) - f_0(x_i) = \mathbb{E}[y_i|x_i, d_i = 1] - \mathbb{E}[y_i|x_i, d_i = 0] \; .$$

The CATE solves the identification problem by comparing the outcomes of treated and non-treated that appear identical in all relevant dimensions, as encoded in x_i.

In contrast to the previously discussed ATE, ATT, ATC and LATEs, the CATE is not a scalar, but a function that can be evaluated at different x_i. When treatment effect heterogeneity is suspected, computing such a CATE-function can be of direct interest. For example, some advertising campaign might attract the young but drive away the old, or a job training program might help those who join voluntarily but have little effect for others. In other cases, one might be forced to compute CATE instead of the unconditional ATEs or LATEs because treatment d_i is not (unconditionally) exogenous and no exogenous IV z_i is available, but one can reasonably argue that treatment (or the IV; see bottom of this section) is exogenous conditional on x_i.[32] In either case, based on CATE, one can compute ATE, ATT and ATC using the LIE, which essentially involves taking expectations of CATE(x_i) either for all observations or only the treated or non-treated ones. For instance, since

$$\mathbb{E}[y_{1i}] = \mathbb{E}[\mathbb{E}[y_{1i}|x_i]] = \mathbb{E}[f_1(x_i)] \quad \text{and} \quad \mathbb{E}[y_{0i}] = \mathbb{E}[\mathbb{E}[y_{0i}|x_i]] = \mathbb{E}[f_0(x_i)] \; ,$$

we have

$$\text{ATE} = \mathbb{E}[\text{CATE}(x_i)] = \int \text{CATE}(x_i) f_{X_i}(x_i) \; ,$$

where the expectation is taken w.r.t. x_i. By analogous arguments, we have

$$\begin{aligned} \text{ATT} &= \mathbb{E}[\text{CATE}(x_i)|d_i = 1] = \int \text{CATE}(x_i) f_{X_i|D_i=1}(x_i) \; , \\ \text{ATC} &= \mathbb{E}[\text{CATE}(x_i)|d_i = 0] = \int \text{CATE}(x_i) f_{X_i|D_i=0}(x_i) \; , \end{aligned}$$

whereby the first expectation is taken w.r.t. $x_i|(d_i = 1)$ and the latter w.r.t. $x_i|(d_i = 0)$.

To estimate CATE, one needs to estimate the functions $f_1(x_i)$ and $f_0(x_i)$. For this purpose, one can use machine learning methods like kernel regression or random forests (see Chapter 14) or matching methods, which are discussed in the following. All such methods that compute CATE based on CIA require, either explicitly or implicitly, the Common Support

[32] In addition, under attrition, conditioning on x_i might render the assumption of random attrition more credible. See Section 13.2.

Assumption (CSA) (or overlap assumption). It states that

$$0 < \mathbb{P}[d_i = 1|x_i] \equiv p(x_i) < 1 \quad \forall\, x_i \in \mathcal{X} \,, \tag{13.7}$$

where $\mathcal{X}$ is the support of x_i and $p(x_i)$ is the conditional treatment probability, the so-called propensity score. Put simply, the CSA requires that for all possible values of x_i there is a chance of observing both treated and non-treated units. If CSA does not hold for some x_i, then CATE(x_i) is not identified at those x_i. For example, if $x_i \in \{0,1\}$ is binary, and all observations with $x_i = 1$ have $d_i = 1$, then we cannot estimate $\mathbb{E}[y_i|x_i = 1, d_i = 0]$ (or at least not non-parametrically; see comparison to the parametric linear regression below).[33]

The idea behind matching methods is to discretize $\mathcal{X}$, the support of x_i, and estimate CATEs for $x_i \in \mathcal{X}_s$ for many different (but finitely many) sets $\mathcal{X}_s$, $s = 1:S$:

$$\text{CATE}(\mathcal{X}_s) = \mathbb{E}[y_i|d_i = 1, x_i \in \mathcal{X}_s] - \mathbb{E}[y_i|d_i = 0, x_i \in \mathcal{X}_s] \,.$$

In other words, we match treated and non-treated units who appear similar enough with regard to their x_i – where we define the degree of similarity by definining the strata $\{\mathcal{X}_s\}_{s=1}^S$ –, and we argue that treatment is as good as random within each such $\mathcal{X}_s$. Corespondingly, the outcomes of non-treated units with $x_i \in \mathcal{X}_s$ are used as counterfactual outcomes for the treated units with $x_i \in \mathcal{X}_s$, and vice versa. As usual, we can estimate this CATE consistently by replacing the expectations by sample averages. Defining $\mathcal{n}_{1,s} = \{i : d_i = 1, x_i \in \mathcal{X}_s\}$ as the treated observations with $x_i \in \mathcal{X}_s$ and similarly for $\mathcal{n}_{0,s}$, with sizes $n_{1,s}$ and $n_{0,s}$, we get

$$\widehat{\text{CATE}}(\mathcal{X}_s) = \frac{1}{n_{1,s}} \sum_{i \in \mathcal{n}_{1,s}} y_i - \frac{1}{n_{0,s}} \sum_{i \in \mathcal{n}_{0,s}} y_i \,.$$

Choosing the granularity of strata $\{\mathcal{X}_s\}_{s=1}^S$ entails a trade-off between credibility and efficiency: under more granular strata, the observations used to construct counterfactual outcomes become more similar and therefore the analysis becomes more credible, but the number of observations in each strata $\mathcal{X}_s$ decreases, generating more noisy estimators. An explicit limit to granularity is set by the common support requirement, as $\widehat{\text{CATE}}(\mathcal{X}_s)$ cannot be computed if there are only treated or only non-treated units in some $\mathcal{X}_s$. This trade-off can also be thought of as a trade-off between generality and efficiency, because common support issues can be remedied when focusing only on a subset of $\mathcal{X}$ with enough treated

[33] Generally, for all methods to compute CATE, to judge whether common support is a problem, one can compute the normalized differences $(\bar{x}_{1j} - \bar{x}_{0j})/\sqrt{\sigma_{1j}^2 + \sigma_{0j}^2}$, where $\bar{x}_{1j} = \frac{1}{n_1}\sum_{i \in \mathcal{n}_1} x_{ij}$ is the sample average of covariate j for the treated observations and $\sigma_{1j}^2 = \frac{1}{n_1}\sum_{i \in \mathcal{n}_1}(x_{ij} - \bar{x}_{1j})^2$ is its variance, and analogously for $\bar{x}_{0j}$ and σ_{0j}^2. The closer this statistic is to zero, the better.

and non-treated in each strata.

Matching methods are similar to RDDs, which match treated and non-treated observations that are close to some threshold that determines treatment status (or the likelihood thereof). However, RDDs construct one such match, whereas matching methods do so for many different strata $\{\mathcal{X}_s\}_{s=1}^S$, yielding heterogeneous treatment effects. Matching methods are also similar to including interaction terms in LRMs. It is easy to see that $\text{CATE}(\mathcal{X}_s)$ coincides with $\beta_{1,s}$ from the linear regression

$$y_i = \sum_{s=1}^S \beta_{0,s}\, \mathbf{1}\left\{x_i \in \mathcal{X}_s\right\} + \sum_{s=1}^S \beta_{1,s} d_i\, \mathbf{1}\left\{x_i \in \mathcal{X}_s\right\} + v_i \tag{13.8}$$

if (i) treatment d_i is indeed exogenous of outcomes y_i conditional on x_i, and if (ii) the strata are granular enough so that v_i is independent of x_i. As before, $\text{CATE}(\mathcal{X}_s) = \beta_{1,s}$ does not imply that the estimators of this object under the non-parametric matching approach and the parametric linear regression model coincide. In this case, this is important insofar as one can compute the OLS estimator $\hat{\beta}_{1,s}$ for all $s = 1 : S$ even if CSA fails for some $\mathcal{X}_s$, whereas computing the above, non-parametric $\widehat{\text{CATE}}(\mathcal{X}_s)$ would be impossible. The reason is that the parametric functional form assumption for y_i in Eq. (13.8) allows us to use all observations when computing $\hat{\beta}_{1,s}$ for each s, whereas, in absence of such an assumption, the non-parametric matching approach uses only observations i with $x_i \in \mathcal{X}_s$ when computing $\widehat{\text{CATE}}(\mathcal{X}_s)$.[34] Nevertheless, even if the linear regression model allows us to avoid addressing common support issues, the resulting estimates $\{\hat{\beta}_{1,s}\}_{s=1}^S$ are only credible if CSA holds.

The CATEs computed by matching are aggregated to the unconditional ATE, ATT and ATC as follows. Suppose that the strata $\{\mathcal{X}_s\}_{s=1}^S$ form a partition of $\mathcal{X}$, the support of x_i. Given $x_i \in \mathcal{X}_s$, let $\widehat{\text{CATE}}_i = \widehat{\text{CATE}}(\mathcal{X}_s)$ be the estimated conditional treatment effect for unit i.[35] We get

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^n \widehat{\text{CATE}}_i\ , \quad \widehat{\text{ATT}} = \frac{1}{n_1}\sum_{i \in \mathcal{n}_1} \widehat{\text{CATE}}_i\ , \quad \text{and} \quad \widehat{\text{ATC}} = \frac{1}{n_0}\sum_{i \in \mathcal{n}_0} \widehat{\text{CATE}}_i\ ,$$

where $\mathcal{n}_1$ and $\mathcal{n}_0$ are the sets of treated and non-treated observations, respectively, with sizes n_1 and n_0. Rather than aggregating the CATEs for all observations or all treated or non-treated observations, one can also aggregate them similarly from a finer to a more coarse level of disaggregation. For example, given CATEs for each age, one can compute CATEs

[34] In such a context, the parametric linear regression model is also referred to as "global", whereas the non-parametric matching approach is "local".

[35] More precisely, we define $\widehat{\text{CATE}}_i = \sum_{s=1}^S \widehat{\text{CATE}}(\mathcal{X}_s)\, \mathbf{1}\left\{x_i \in \mathcal{X}_s\right\}$.

for 20-30 year-olds, 30-40 year-olds, etc.

Such non-parametric matching methods become increasingly infeasible when the number of observations available is limited relative to the dimension of the vector of covariates x_i on which the CATEs are conditioned. Machine learning methods are somewhat better equipped to address the curse of dimensionality, but their approach for conditioning on x_i when computing CATEs is rather intransparent. One possible solution is the linear regression model, which, however, due to its parametric assumptions, might be of limited credibility (see above). Another possible solution are semi-parametric matching methods, which are based on preliminary, typically parametric estimates of the propensity scores $p(x_i)$. They are based on the fact that independence of treatment d_i and outcomes y_i conditional on x_i implies independence conditional on $p(x_i)$.

Under Inverse Propensity Score Weighting (IPW), an estimator for ATE based on the one-dimensional $p(x_i)$ rather than the high-dimensional x_i is obtained as follows. The Appendix shows that we can write $f_1(x_i)$ and $f_0(x_i)$ from above as

$$f_1(x_i) = \mathbb{E}\left[\frac{d_i y_i}{p(x_i)} | x_i\right] \quad \text{and} \quad f_0(x_i) = \mathbb{E}\left[\frac{(1-d_i)y_i}{1-p(x_i)} | x_i\right] \ . \tag{13.9}$$

In turn, we get

$$\begin{aligned} \text{ATE} = \mathbb{E}[\text{CATE}(x_i)] &= \mathbb{E}\left[\mathbb{E}\left[\frac{d_i y_i}{p(x_i)} - \frac{(1-d_i)y_i}{1-p(x_i)} \mid x_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{[d_i - p(x_i)]y_i}{p(x_i)[1-p(x_i)]} \mid x_i\right]\right] = \mathbb{E}\left[\frac{[d_i - p(x_i)]y_i}{p(x_i)[1-p(x_i)]}\right] \ , \end{aligned}$$

which allows us to construct the estimator

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{[d_i - \hat{p}(x_i)]y_i}{\hat{p}(x_i)[1-\hat{p}(x_i)]} \ .$$

Note that the estimators become noisy if there are many $\hat{p}(x_i)$ close to zero or close to one. Under IPW with common support issues, the trade-off is between generality and efficiency, as one can obtain a less noisy estimator by limiting the analysis to a subset of $\mathcal{X}$ with non-extreme values of $\hat{p}(x_i)$.

Under Pair-Matching (PM), we estimate the ATT by comparing the outcome of each treated unit to the outcome of the non-treated unit that is closest in terms of propensity score:

$$\widehat{\text{ATT}}_{PM} = \frac{1}{n_1}\sum_{i \in n_1}^{n} \left\{ y_i - \sum_{j \in n_1} \mathbf{1}\left\{ j = \arg\min_j |\hat{p}(x_j) - \hat{p}(x_i)| \right\} y_j \right\} \ .$$

Thereby, one can also use other distance measures, like the Mahalanobis distance $(x_j - x_i)'\hat{\Sigma}_x^{-1}(x_j - x_i)$, where $\hat{\Sigma}_x$ is the sample covariance matrix of the covariates. Under Radius Matching (RM), instead of using the outcome of the single nearest non-treated unit, one uses the average outcomes of M nearest non-treated units, which reduces the variance of the estimator. To deal with common support issues, one can only consider treated units for which reasonably close control units can be found and/or one can repeat the analysis deleting the control units which appear in too many comparisons.[36]

While this section discussed approaches to compute conditional average treatment effects based on the assumption of independence of potential outcomes (y_{1i}, y_{0i}) and treatment d_i conditional on x_i, analogous approaches can be implemented under conditional independence of potential outcomes and an IV z_i. Again, they can be motivated by a desire to estimate heteroeneity in treatment effects or to render the analysis more credible, as sometimes covariates x_i need to be included to shut down effects of z_i on y_i that do not operate via d_i, but x_i. Analogously to above, we get

$$\text{LATE}(\mathcal{X}_s) = \frac{\mathbb{E}[y_i|z_i = 1, x_i \in \mathcal{X}_s] - \mathbb{E}[y_i|z_i = 0, x_i \in \mathcal{X}_s]}{\mathbb{E}[d_i|z_i = 1, x_i \in \mathcal{X}_s] - \mathbb{E}[d_i|z_i = 0, x_i \in \mathcal{X}_s]} \, .$$

Also, semi-parametric methods can be applied by compressing the information in x_i into $\mathbb{P}[z_i = 1|x_i]$.

13.3.4 Difference-in-Differences

Difference-in-Differences (DiD) estimation can be applied if we observe treated and non-treated units in two periods, before and after treatment. Let $y_{0i,t=1}$ be the potential outcome in absence of treatment at $t = 1$, before treatment occurred, and let $y_{0i,t=2}$ and $y_{1i,t=2}$ be the potential outcomes without and with treatment, respectively, at $t = 2$, after treatment occurred. We observe the following outcomes at $t = 1$ and $t = 2$:

$$y_{i,t=1} = y_{0i,t=1} \; \forall \, i \quad \text{and} \quad y_{i,t=2} = \begin{cases} y_{0i,t=2} & \text{for } \; i \in \mathcal{N}_0 \equiv \{i : d_i = 0\} \\ y_{1i,t=2} & \text{for } \; i \in \mathcal{N}_1 \equiv \{i : d_i = 1\} \end{cases} .$$

Note that this statement rules out anticipation effects; for every treated unit i, the outcome observed at $t = 1$, $y_{i,t=1}$, is indeed equal to $y_{0i,t=1}$. In other words, we rule out the possibility that some units were treated already in $t = 1$.

[36] Such approaches to adjust matching methods in light of common support issues are referred to as "trimming".

We are interested in

$$\text{ATT} = \mathbb{E}[y_{1i,t=2} - y_{0i,t=2} | d_i = 1] \ .$$

Without further assumptioons, this quantity is not estimable because we do not observe $y_{0i,t=2}$ for units with $d_i = 1$. Comparing instead the outcomes of treated units before and after treatment,

$$\mathbb{E}[y_{i,t=2} - y_{i,t=1} | d_i = 1] = \mathbb{E}[y_{1i,t=2} - y_{0i,t=1} | d_i = 1] \ ,$$

does not reveal the ATT because aspects of the environment other than the treatment typically change over time and affect outcomes, i.e. $\mathbb{E}[y_{0i,t=1}|d_i = 1] \neq \mathbb{E}[y_{0i,t=2}|d_i = 1]$. Comparing the outcomes of treated and non-treated units after treatment,

$$\mathbb{E}[y_{i,t=2}|d_i = 1] - \mathbb{E}[y_{i,t=2}|d_i = 0] = \mathbb{E}[y_{1i,t=2}|d_i = 1] - \mathbb{E}[y_{0i,t=2}|d_i = 0] \ ,$$

does not reveal the ATT either because treated and non-treated units typically differ in their potential outcomes, i.e. $\mathbb{E}[y_{0i,t=2}|d_i = 0] \neq \mathbb{E}[y_{0i,t=2}|d_i = 1]$.

Under the assumption that, in absence of treatment, treated and non-treated (on average) would have been subject to the same time trend, we can use the observed change in outcomes from $t = 1$ to $t = 2$ for the control group to infer the (average) hypothetical outcomes for the treatment group that would have been obtained in $t = 2$ in absence of treatment. Formally, this Common Trend Assumption (CTA) is

$$\mathbb{E}\left[y_{0i,t=2}|d_i = 1\right] - \mathbb{E}\left[y_{0i,t=1}|d_i = 1\right] = \mathbb{E}\left[y_{0i,t=2}|d_i = 0\right] - \mathbb{E}\left[y_{0i,t=1}|d_i = 0\right] \ .$$[37]

Under CTA, we can write the above ATT in terms of observables as a difference-in-differences, comparing the observed change in average outcomes for treated and controls:

$$\begin{aligned}
\text{ATT} &= \mathbb{E}[y_{1i,t=2}|d_i = 1] - \mathbb{E}[y_{0i,t=2}|d_i = 1] \\
&= \mathbb{E}[y_{1i,t=2}|d_i = 1] - \mathbb{E}[y_{0i,t=1}|d_i = 1] - \{\mathbb{E}[y_{0i,t=2}|d_i = 1] - \mathbb{E}[y_{0i,t=1}|d_i = 1]\} \\
&= \{\mathbb{E}[y_{1i,t=2}|d_i = 1] - \mathbb{E}[y_{0i,t=1}|d_i = 1]\} - \{\mathbb{E}[y_{0i,t=2}|d_i = 0] - \mathbb{E}[y_{0i,t=1}|d_i = 0]\} \\
&= \{\mathbb{E}[y_{i,t=2}|d_i = 1] - \mathbb{E}[y_{i,t=1}|d_i = 1]\} - \{\mathbb{E}[y_{i,t=2}|d_i = 0] - \mathbb{E}[y_{i,t=1}|d_i = 0]\} \ .
\end{aligned}$$

[37] By rearranging terms, we can also state it as a "bias stability assumption":

$$\mathbb{E}\left[y_{0i,t=2}|d_i = 1\right] - \mathbb{E}\left[y_{0i,t=2}|d_i = 0\right] = \mathbb{E}\left[y_{0i,t=1}|d_i = 1\right] - \mathbb{E}\left[y_{0i,t=1}|d_i = 0\right] \ ,$$

i.e. if we compared treated and controls in $t = 2$ in absence of treatment, we would get the same as when comparing them in $t = 1$, when treatment actually did not occur (yet).

In turn, we can estimate it as

$$\widehat{\text{ATT}} = \left\{ \frac{1}{n_1} \sum_{i \in n_1} y_{i,t=2} - \frac{1}{n_1} \sum_{i \in n_1} y_{i,t=1} \right\} - \left\{ \frac{1}{n_0} \sum_{i \in n_0} y_{i,t=2} - \frac{1}{n_0} \sum_{i \in n_0} y_{i,t=1} \right\} .^{38} \tag{13.10}$$

An example of a study applying the DiD-estimator of the ATT is Card and Krueger (1994). They estimate the effect of an increase in minimum wages in New Jersey in 1992 on employment by comparing the change in employment across fast-food restaurants in New Jersey to the corresponding change in neighboring Pennsylvania, whose minimum wage remained unchanged.

The CTA is not testable. To convince an audience that it holds, one can show that, before treatment, average outcomes for treated and controls changed in a similar way over time. An even more convincing case for CTA can be made if one estimates the ATT for two periods in which no treatment occurred and finds an effect not significantly different from zero. However, for both of these arguments, one needs data for at least two pre-treatment periods. In absence of such data, and at the very least, one should prove that background characteristics between treated and non-treated units are balanced (pre-treatment).[39] CTA is likely to be violated if treatment d_i is assigned based on pre-treatment outcomes, $y_{i,t=1} = y_{0i,t=1}$. Also, note that CTA is sensitive to non-linear transformations of outcomes; a common (linear) trend for log y_i implies that there is no common (linear) trend for y_i.

The DiD-ATT above coincides with the parameter β_3 from the linear regression

$$y_{it} = \beta_0 + \beta_1 d_i + \beta_2 \, \mathbf{1}\{t=2\} + \beta_3 d_i \, \mathbf{1}\{t=2\} + u_{it} \; . \tag{13.11}$$

However, to give this parameter a causal interpretation, CTA needs to hold and we need to rule out anticipation effects. In the linear regression, CTA manifests itself as follows:

$$\begin{aligned} & \mathbb{E}\left[y_{0i,t=2} - y_{0i,t=1} | d_i = 1\right] = \mathbb{E}\left[y_{0i,t=2} - y_{0i,t=1} | d_i = 0\right] \\ \Leftrightarrow \quad & \mathbb{E}\left[\beta_0 + \beta_2 + u_{0i,t=2} - \beta_0 - u_{0i,t=1} | d_i = 1\right] = \mathbb{E}\left[\beta_0 + \beta_2 + u_{0i,t=2} - \beta_0 - u_{0i,t=1} | d_i = 0\right] \\ \Leftrightarrow \quad & \mathbb{E}\left[u_{0i,t=2} - u_{0i,t=1} | d_i = 1\right] = \mathbb{E}\left[u_{0i,t=2} - u_{0i,t=1} | d_i = 0\right] \; . \end{aligned}$$

Note that this assumption is not part of the typical assumptions made under OLS estimation of linear regressions.[40] As in Section 13.3.3, even though β_3 coincides with the ATT above,

[39] There are more such checks that can be done. For example, to provide evidence for common trends in employment in fast-food restaurants in New Jersey and Pennsylvania, Card and Krueger (1994) show that, over the same two periods, the change in employment in restaurants offering higher wages is not significantly different in the two states.

[40] The DiD-ATT also coincides with β_3 if unit-fixed effects α_i are added to Eq. (13.11). It turns out that,

the parametric OLS estimator of β_3 does not coincide with the non-parametric DiD estimator in Eq. (13.10). In this context, this is important insofar as OLS estimation requires a balanced panel dataset, whereas to implement the non-parametric DiD estimation, repeated crosss-sections suffice. One can allow for different sets of treated and non-treated units in $t = 1$ and $t = 2$ in Eq. (13.10) by defining $n_{1,t=1}$, $n_{1,t=2}$, $n_{0,t=1}$ and $n_{0,t=2}$. Standard errors for either estimator need to take into account that $y_{i,t=2}$ and $y_{i,t=1}$ (in Eq. (13.10)) and $u_{i,t=2}$ and $u_{i,t=1}$ are not independent, leading to "clustered" standard errors (clustered at the level of i).

Analogous extensions as in Section 13.3.3 can be made. By stating the CTA conditional on some covariates x_i,

$$\mathbb{E}\left[y_{0i,t=2} - y_{0i,t=1} | x_i, d_i = 1\right] = \mathbb{E}\left[y_{0i,t=2} - y_{0i,t=1} | d_i, d_i = 0\right] ,$$

we obtain a conditional ATT (CATT) that takes the form of a difference-in-differences:

$$\begin{aligned} \text{CATT}(x_i) &= \mathbb{E}[y_{1i,t=2} | x_i, d_i = 1] - \mathbb{E}[y_{0i,t=2} | x_i, d_i = 1] \\ &= \mathbb{E}[y_{i,t=2} - y_{i,t=1} | x_i, d_i = 1] - \mathbb{E}[y_{i,t=2} - y_{i,t=1} | x_i, d_i = 0] . \end{aligned}$$

It can be aggregated to obtain the unconditional ATT using the LIE: ATT $= \mathbb{E}[\text{CATT}(x_i)] = \int \text{CATT}(x_i) f_{X_i|D_i=1}(x_i) dx_i$. The CATT can be estimated by dividing $\mathcal{X}$, the sample space of x_i, into strata $\{\mathcal{X}_s\}_{s=1}^{S}$ and considering the sample analogue of

$$\text{CATT}(\mathcal{X}_s) = \mathbb{E}[y_{i,t=2} - y_{i,t=1} | d_i = 1, x_i \in \mathcal{X}_s] - \mathbb{E}[y_{i,t=2} - y_{i,t=1} | d_i = 0, x_i \in \mathcal{X}_s] ,$$

given by

$$\widehat{\text{CATT}}(\mathcal{X}_s) = \left\{ \frac{1}{n_{1,s}} \sum_{i \in n_{1,s}} y_{i,t=2} - \frac{1}{n_{1,s}} \sum_{i \in n_{1,s}} y_{i,t=1} \right\} - \left\{ \frac{1}{n_{0,s}} \sum_{i \in n_{0,s}} y_{i,t=2} - \frac{1}{n_{0,s}} \sum_{i \in n_{0,s}} y_{i,t=1} \right\} .$$

These CATTs are once again equivalent to the corresponding parameters in a linear regression

in this simple setting, these are irrelevant when interest lies in estimating the ATT. The reason is that the ATT considers average treatment effects, which means that any such unit-fixed effects are differenced out. What stays is the group-fixed effect $\beta_1 d_i$, capturing pre-treatment differences in average outcomes across treated and controls.

as in Eq. (13.11), where each term (but the error) is interacted with dummies $\mathbf{1}\{x_i \in \mathcal{X}_s\}$:

$$\begin{aligned} y_{it} =& \sum_{s=1}^{S} \beta_{0,s}\, \mathbf{1}\{x_i \in \mathcal{X}_s\} + \sum_{s=1}^{S} \beta_{1,s} d_i\, \mathbf{1}\{x_i \in \mathcal{X}_s\} \\ &+ \sum_{s=1}^{S} \beta_{2,s}\, \mathbf{1}\{t=2\}\, \mathbf{1}\{x_i \in \mathcal{X}_s\} + \sum_{s=1}^{S} \beta_{3,s} d_i\, \mathbf{1}\{t=2\}\, \mathbf{1}\{x_i \in \mathcal{X}_s\} + u_{it}\ . \end{aligned} \tag{13.12}$$

Computing CATTs can be of interest because one suspects separate time trends for, say, men and women, across education levels or across different regions in a country, meaning that CTA holds only conditional on x_i. In that sense, x_i should contain all potential sources of group-specific trends. However, even if the unconditional CTA were not violated, one might compute CATTs if interest lies in treatment heterogeneity. As before, this analysis requires a CSA, and choosing the level of granularity of $\{\mathcal{X}_s\}_{s=1}^{S}$ involves a trade-off between credibility on the one hand and efficiency and generality on the other.

Sometimes, potential sources of group-specific trends are included in Eq. (13.11) differently than in Eq. (13.12). For example, given a single (scalar) background characteristic x_i, one might estimate the models

$$y_{it} = \beta_0 + \gamma_0 x_i + \beta_1 d_i + \beta_2\, \mathbf{1}\{t=2\} + \beta_3 d_i\, \mathbf{1}\{t=2\}\ ,$$

or

$$\begin{aligned} y_{it} =& \beta_0 + \beta_1 d_i + \beta_2\, \mathbf{1}\{t=2\} + \beta_3 d_i\, \mathbf{1}\{t=2\} \\ &+ \gamma_0 x_i + \gamma_1 d_i x_i + \gamma_2\, \mathbf{1}\{t=2\}\, x_i + \gamma_3 d_i\, \mathbf{1}\{t=2\}\, x_i + u_{it}\ , \end{aligned}$$

Such ad-hoc inclusion of covariates can be fine from a modeling perspective, but the parameters β_3 and γ_3 do not (necessarily) have a causal interpretation![41] They can be used to estimate treatment effects through the lens of the specified model, e.g. to decompose the overall correlation of treatment and outcomes into partial correlations by accounting for particular, parametric differences of the correlation of treatment with outcomes for different x_i as well as differences in the correlation of outcomes and x_i over time and across the treatment and control groups. In other words, the parameters β_3 and γ_3 only have a causal interpretation if the model is correctly specified or can be motivated in a non-parametric way, from a causal inference point of view. There is a surge of papers in recent years that illustrate that such specifications do not have causal interpretations and propose alternative, typically non-parametric estimation approaches that do. The same holds for extensions

[41] Except if $x_i \in \{0, 1\}$ is binary, in which case this specification is equivalent to Eq. (13.12).

of Eq. (13.11) to multiple periods over which treatment occurs ("staggered treatment") or when conducting an event study around a single treatment period. In other words, β_3 from Eq. (13.11) does not (necessarily) retain its causal interpretation if that model is estimated using observations for more than one period over which treatment occurs or if terms like $\sum_{l=-H}^{H} \gamma_l \; \mathbf{1}\{t = 2 + l\} + \sum_{l=H}^{H} \delta_l d_i \; \mathbf{1}\{t = 2 + l\}$ are added, capturing the change in average outcomes for the control and treatment groups, respectively, from H periods before treatment to H periods after treatment.

A potential problem in DiD-estimation is when the composition of the two treatment and control groups changes in response to treatment, i.e. if some units i move from the control- ($d_i = 0$) to the treatment-group ($d_i = 1$) or vice versa. In the context of Card and Krueger (1994), this might happen if between the two periods of observation, some workers moved from Pennsylvania to New Jersey, towards the state with the increase in minimum wages, or vice versa, away from the state with the increase in minimum wages. From a causal inference perspective, the former are always-takers, the latter are never-takers, and the analysis can be extended analogously as done in Section 13.2, using location in $t = 1$ as an IV for location in $t = 2$ and computing a local ATT for the workers who did not move.

There are numerous ways along which DiD-approaches can be (and are) extended. One is staggered treatment, i.e. treatment occurring for different units at different periods, whereby treatment may or may not be an absorbing state (meaning that units once treated stay treated in future periods). Another one is treatment occurring over several periods (for the same unit). Some DiD-approaches exploit the fact that treatment occurs typically along more than just two dimensions, groups $d_i \in \{0, 1\}$ and time $t \in \{1, 2\}$. For example, if treatment occurs at some period in some states and for some age-groups, but not in other states and/or for other age-groups, one might get more precise DiD estimators by exploiting both the state- and age-dimensions rather than just one of them (comparing treated and non-treated of the same age across states or in the same state across ages). Taking into account that several different treatments may occur around the same time and recalling the potential to condition DiD-analyses on background characteristics or to incorporate imperfect compliance with IVs, this yields a plethora of different DiD-type methods!

Appendix

Finding Causality with Experimental Data

On the Equivalence of LATE and 2SLS The first-stage regression is

$$d_i = \alpha_0 + \alpha_1 z_i + u_i \ ,$$

where $\alpha_1 = \mathbb{E}[d_i|z_i = 1] - \mathbb{E}[d_i|z_i = 0]$ is given by $\alpha_1 = \text{Cov}(z_i, d_i)/\mathbb{V}[z_i]$. Under independent treatment assignment z_i, $\text{Cov}(z_i, d_i) = \text{Cov}(z_i, \alpha_0 + \alpha_1 z_i + u_i) = \alpha_1 \mathbb{V}[z_i]$. The second-stage regression is

$$y_i = \beta_0 + \beta_1 \hat{d}_i + e_i \ ,$$

where $\hat{d}_i = \alpha_0 + \alpha_1 z_i$, and we can write

$$y_i = \gamma_0 + \gamma_1 z_i + e_i$$

with $\gamma_0 = \beta_0 + \beta_1 \alpha_1$ and $\gamma_1 = \beta_1 \alpha_1$. The LATE is given by $\beta_1 = \gamma_1/\alpha_1 = \text{Cov}(z_i, y_i)/\text{Cov}(z_i, d_i)$, i.e. we divide the ITT $\gamma_1 = \mathbb{E}[y_i|z_i = 1] - \mathbb{E}[y_i|z_i = 0]$ by $\alpha_1 = \mathbb{E}[d_i|z_i = 1] - \mathbb{E}[d_i|z_i = 0]$, the effect of z_i on d_i from the first stage. Provided that there are at least some compliers, instrument relevance is sastisfied: $\text{Cov}(z_i, d_i) \neq 0$. Also, if treatment-assignment is indeed randomized, instrument exogeneity is satisfied: $\text{Cov}(z_i, e_i) = 0$. Again, covariates x_i might be added to both the first- and second-stage regressions in order to reduce the variance of the estimator for β_1,[42] though more work needs to be done to establish a causal interpretation of β_1 in that case (see Section 13.3.3).

[42] The same covariates need to be added to both regressions. See Section 7.4.

Finding Causality with Observational Data

Derivation of result in Eq. (13.9) Note that $f_1(x_i) \equiv \mathbb{E}[y_{1i}|x_i] = \mathbb{E}[y_{1i}|x_i, d_i]$ and $p(x_i) \equiv \mathbb{E}[d_i|x_i]$. Using these definitions and results, we have

$$\begin{aligned}
f_1(x_i) &= \mathbb{E}\left[\frac{d_i}{p(x_i)} \mid x_i\right] f_1(x_i) \\
&= \mathbb{E}\left[\frac{d_i \mathbb{E}[y_{1i} \mid x_i]}{p(x_i)} \mid x_i\right] \\
&= \mathbb{E}\left[\frac{d_i \mathbb{E}[y_{1i} \mid x_i, d_i]}{p(x_i)} \mid x_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{d_i y_{1i}}{p(x_i)} \mid x_i, d_i\right] \mid x_i\right] \\
&= \mathbb{E}\left[\frac{d_i y_{1i}}{p(x_i)} \mid x_i\right] \\
&= \mathbb{E}\left[\frac{d_i y_i}{p(x_i)} \mid x_i\right] \ .
\end{aligned}$$

Analogously, one can show that $f_0(x_i) = \mathbb{E}\left[\frac{(1-d_i)y_i}{1-p(x_i)} \mid x_i\right]$.