12 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 12.1 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 12.2 and using quasi-experimental methods in Section 12.3. 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)).

12.1 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 8.2).

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, $y_i = y_{0i}$ holds, whereas for the non-treated subpopulation, $y_i = y_{1i}$ 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 12.2 and 12.3) –, 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:

$$\frac{1}{n_1} \sum_{i \in n_1} y_i - \frac{1}{n_0} \sum_{i \in n_0} y_i \quad \xrightarrow{p} \quad \mathbb{E}[y_{1i}|d_i = 1] - \mathbb{E}[y_{0i}|d_i = 0] = \text{ATE} = \text{ATT} = \text{ATC} \, . \tag{12.1}$$

This environment and attempts of creating it are discussed in Section 12.2.

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,

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

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

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 12.3.

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 3.4). Consider the regression with an intercept and d_i as the only covariate:

$$y_i=\beta_0+\beta_1 d_i+u_i\ , \tag{12.2}$$

[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

$$\begin{aligned}\mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0] &= \mathbb{E}[\beta_0 + \beta_1 d_i + u_i|d_i = 1] - \mathbb{E}[\beta_0 + \beta_1 d_i + u_i|d_i = 0] \\ &= \beta_1 + \mathbb{E}[u_i|d_i = 1] - \mathbb{E}[u_i|d_i = 0] \ , \end{aligned} \tag{12.3}$$

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 3.4, 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.

12.2 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. (12.1), 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. (12.2).[14] This is not necessarily true in other settings than the present one with randomized, binary treatment.[15]

Sometimes, covariates are added to that regression in Eq. (12.2) 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. 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 3.5 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 12.1): 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 ($d_i = 1 - z_i$).[16] Typically, it is assumed that there are no defiers, which is referred to as the monotonicity assumption.[17]

[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] 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.

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

Table 12.1: 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), 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.[18] [19] We can correct for this downward bias by scaling up the ITT by the fraction of compliers in our sample, $\mathbb{E}[d_i|z_i = 1] - \mathbb{E}[d_i|z_i = 0] = \mathbb{P}[d_i = 1|z_i = 1] - \mathbb{P}[d_i = 1|z_i = 0]$.[20] 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:

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

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 the parametric two-stages least squares (2SLS) Instrument Variable (IV) estimator, in which

[18] 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.

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

[20] 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.

treatment-assignment z_i is used as an IV for treatment-status d_i. [21] 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).[22] 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 12.3.3) and thereby ignore attrition.[23] 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 variables, one could use the lowest and highest observed values of the outcome variable to

[21] 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 $\hat{d}_i = \alpha_0 + \alpha_1 z_i$.

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

[23] 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.