Chapter 6: Policy Evaluation Methods: Treatment Effects

Joan Llull

Microeconometrics
IDEA PhD Program

I. Potential Outcomes and Causality

A. *Potential Outcomes, Selection Bias, and Treatment Effects*

Consider the population of individuals that are susceptible of a treatment. Let Y_{1i} denote the outcome for an individual i if exposed to the treatment ($D_i = 1$), and let Y_{0i} be the outcome for the same individual if not exposed ($D_i = 0$). The *treatment effect* for individual i is thus $Y_{1i} - Y_{0i}$. Note that Y_{1i} and Y_{0i} are *potential outcomes* in the sense that we only observe one of the two:

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i). \tag{1}$$

This poses the main challenge of this approach, as the treatment effect can not be computed for a given individual. Fortunately, our interest is not in treatment effects for specific individuals *per se*, but, instead, in some characteristics of their distribution, like some average.

We mainly focus on two parameters of interest. The first one is the *average treatment effect* (ATE):

$$\alpha_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}], \tag{2}$$

and the second one is *average treatment effect on the treated* (TT):

$$\alpha_{TT} \equiv \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]. \tag{3}$$

Note the subtle difference between the two. The first one is an ideal parameter of interest, but difficult to obtain: the average of the treatment effects for the population. The second one, which is often easier to obtain, is the average treatment effect computed over treated individuals, that is, for the individuals that actually experiment treatment.

The reason why the second parameter is easier to identify is, precisely, that we only observe Y_i. Let β denote the difference in mean outcomes for treated and

untreated individuals, which can be rewritten as:

$$\begin{aligned}\beta &\equiv \mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0] \\ &= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i=1]}_{\alpha_{TT}} + \underbrace{(\mathbb{E}[Y_{0i}|D_i=1] - \mathbb{E}[Y_{0i}|D_i=0])}_{\text{selection bias}}. \end{aligned} \tag{4}$$

The second term, which we call "selection bias" indicate the difference in untreated potential outcomes between treated and untreated individuals. A nonzero bias may result from a situation in which treatment status is the result of individual decisions where those with low Y_0 choose treatment more frequently than those with high Y_0 or vice versa. In the hospital example, treated individuals (those hospitalized) would also be less healthy had they been at home (negative bias). Thus, the comparison of health between inpatients and other individuals gives a negatively biased estimate of the effect of hospitalization.

From a structural model of D_i and Y_i, one could obtain the implied average treatment effects. Here, they are instead defined with respect to the distribution of potential outcomes, so that, relative to the structure, they are *reduced-form causal effects*. Econometrics has conventionally distinguished between reduced form effects, uninterpretable but useful for prediction, and structural effects, associated with rules of behavior. The treatment effects provide this intermediate category between predictive and structural effects, in the sense that recovered parameters are causal effects, but, as reduced form effects, they are uninterpretable outside of the sample/population of interest and the treatment implemented (or, in other words, they lack external validity). Furthermore, an important assumption of the potential outcome representation is that the effect of the treatment on one individual is independent of the treatment received by other individuals. This excludes equilibrium or feedback effects, as well as strategic interactions among agents. Hence, the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

The sample average version of β is given by:

$$\begin{aligned}\beta^S &\equiv \bar{Y}_T - \bar{Y}_C \\ &\equiv \frac{1}{N_1}\sum_{\text{i=1}}^{\text{N}} Y_i D_i - \frac{1}{N_0}\sum_{\text{i=1}}^{\text{N}}(1-D_i)Y_i, \end{aligned} \tag{5}$$

where $N_0 \equiv N - N_1$ is the number of untreated (or control) individuals.

B. *Identification of Treatment Effects under Different Assumptions*

The identification of the treatment effects depends on the assumptions we make on the relation between potential outcomes and the treatment. The easiest case is when the distribution of the potential outcomes is independent of the treatment:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i. \tag{6}$$

This situation is typical in randomized experiments, where individuals are assigned to treatment or control in a random manner. For example, this occurs, for a given school, in the random assignment of pupils to different class sizes implemented in a randomized experiment called STAR that we will discuss as an example in next chapter. When this is the case, $F(Y_{1i}|D_i = 1) = F(Y_{1i})$, and $F(Y_{0i}|D_i = 0) = F(Y_{0i})$, which implies that $\mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_i|D_i = 1]$ and $\mathbb{E}[Y_{0i}] = \mathbb{E}[Y_{0i}|D_i = 0] = \mathbb{E}[Y_i|D_i = 0]$, and, as a result, $\alpha_{ATE} = \alpha_{TT} = \beta$. Thus, an unbiased estimate of α_{ATE} is given by the difference between average outcomes of treated and control individuals:

$$\widehat{\alpha}_{ATE} = \bar{Y}_T - \bar{Y}_C = \beta^S. \tag{7}$$

In this context, there is no need to "control" for other covariates, unless there is direct interest in their marginal effects, or we want to compute effects for specific groups (we return to this point below).

A less restrictive assumption is *conditional independence*:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|X_i, \tag{8}$$

where X_i is a vector of covariates. This situation is known as matching, as for each "type" of individual (i.e. each value of covariates) we can match treated and control individuals, so that the latter act as counterfactuals for the former. Conditional independence implies that the above results are valid for a given X_i, that is $\mathbb{E}[Y_{1i}|X_i] = \mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_i|D_i = 1, X_i]$ and $\mathbb{E}[Y_{0i}|X_i] = \mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_i|D_i = 0, X_i]$, and, as a result:

$$\begin{aligned}
\alpha_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}] &= \mathbb{E}[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]] \\
&= \int \mathbb{E}[Y_{1i} - Y_{0i}|X_i] dF(X_i) \\
&= \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i).
\end{aligned} \tag{9}$$

In words, the bottom expression computes the difference in average observed outcomes of treated and control individuals that share each value of X_i, and integrate

over the distribution of X_i. Thus, it is "matching" treated individuals with controls that share the same X_i. Similarly, the treatment effect on the treated is:

$$\begin{aligned}\alpha_{TT} &= \int \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1, X_i]dF(X_i|D_i = 1) \\ &= \int \mathbb{E}[Y_i - \mathbb{E}[Y_{0i}|D_i = 1, X_i]|D_i = 1, X_i]dF(X_i|D_i = 1) \\ &= \int \mathbb{E}[Y_i - \mu_0(X_i)|D_i = 1, X_i]dF(X_i|D_i = 1), \end{aligned} \tag{10}$$

where $\mu_0(X_i) \equiv \mathbb{E}[Y_i|D_i = 0, X_i]$, and we use the fact that $\mathbb{E}[Y_i|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|X_i] = \mathbb{E}[Y_{0i}|D_i = 1, X_i]$. The function $\mu_0(X_i)$ is used as an imputation device (matching) for Y_{0i}.

Finally, sometimes we cannot assume conditional independence:

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i|X_i. \tag{11}$$

In this case, we will need some variable Z_i that provides *exogenous variation* in the treatment, meaning that it satisfies the independence assumption:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i|X_i, \tag{12}$$

and the relevance condition:

$$Z_i \not\perp\!\!\!\perp D_i|X_i. \tag{13}$$

As we discuss below, in this context we are only going to be able to identify an average treatment effect for a subgroup of individuals, and we call the resulting parameter a *local average treatment effect*.

II. Randomized Control Trials and Natural Experiments

A. Random Assignment and Treatment Effects

In the treatment effect approach, a *randomized field trial* is regarded as the ideal research design. Observational studies are seen as more speculative attempts to generate the force of evidence of experiments.

There is a long history of randomized field trials in social welfare in the U.S., beginning in the 1960s (see Moffitt, 2003, for a review). Early experiments had many flaws due to the lack of experience in designing them, and in data analysis. During the 1980s, the U.S. Federal Government started to encourage states to use experimentation, eventually becoming almost mandatory. The analysis of the 1980s experimental data consisted of simple treatment-control differences. The

force of the results had a major influence on the 1988 legislation. In spite of these developments, randomization encountered resistance from many U.S. states on ethical grounds. Even more so in other countries, where treatment groups have often been formed by selecting areas for treatment instead of individuals.

Experiments are often very expensive, and often difficult to implement. However, nature sometimes do the job, providing *natural experiments*. Very illustrative to this end is the way in which science connected cholera and the quality of drinking water in the SoHo in London, in 1854. In the 19th century, London suffered from periodic cholera epidemics in which many died. Cholera was believed to be caused by bad air quality, but John Snow (a medical doctor) suspected that instead it was caused by bad water quality (though he had no theory of why). In order to use experimental data to test this hypothesis, one could randomly give some people good water and some people bad water. However, there are good ethical reasons why this experiment cannot be implemented on people.

In 1854 there was a severe outbreak of cholera in Soho. Snow thought contamination of the pump in Broad Street was the source of the problem. He found those for whom this was the closest pump were more likely to die, but in nearby workhouse fewer people died (they had their own well). The brewery on Broad Street itself reported no deaths (they also had their own well —though the men normally only drank beer). These two groups breathed the same air but had access to different water. A further piece of evidence was two isolated deaths, one in Hampstead, one in Islington, of an aunt and her niece. The aunt was in the habit of having a barrel of water delivered from the Broad Street pump every day (she liked the taste) and the niece had paid her a visit. Thus, even though this variation in who drank what water was not assigned at random by any researcher, he built a powerful case that "bad water" was the source of problem. This accidental variation in the source of water can be considered "as good as randomly assigned", because of different water sources across houses in the same street and sometimes even across apartments within houses. Some houses got their drinking water supply from companies such as Lambeth that sourced upstream, i.e. above sewage discharge points, while other houses were supplied by companies such as Southwark and Vauxhall that sourced downstream, i.e. from dirtier water. Snow identified the water companies for the houses with cholera deaths as well as the total number of houses served by each company in his study area. And results corroborated his suspicion.

In a controlled experiment, treatment status is randomly assigned by the re-

searcher, which by construction, ensures independence:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i. \tag{14}$$

As noted above, this eliminates the selection bias (and implies $\alpha_{TT} = \beta$), as:

$$\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0] = \mathbb{E}[Y_{0i}]. \tag{15}$$

It also implies $\alpha_{ATE} = \alpha_{TT} = \beta$, as $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i}]$. Thus, the average treatment effect can be estimated by a simple linear regression of the observed outcome Y_i on the treatment dummy D_i and a constant.

B. Introduction of Additional Regressors

The discussion above shows that econometrics would be very easy if all data was from (well executed) randomized control experiments: one could get causal effects simply by comparing means, there would be no need for matrix algebra or even multiple regressions, and no need to collect any variables other than treatment status and the outcome variable.

However, even in this setup, there are situations in which additional regression can be useful. Let W_i denote a vector of additional possible regressors. Randomization ensures consistency, even if they are not included. The omitted variable bias formula is:

$$\gamma \frac{\text{Cov}(W_i, D_i)}{\text{Var}(D_i)}, \tag{16}$$

which is equal to zero because randomization implies $\text{Cov}(W_i, D_i) = 0$. This is so unless W_i is a "bad control", that is, an intermediate outcome that is affected by the treatment.

One advantage of including additional controls is that, if they are relevant, this would typically increase precision in the estimated average treatment effect. Intuitively this is so because by holding constant other characteristics that affect the outcomes, we are reducing the variance of U_i. More formally, one can apply the partial regression results by Frisch and Waugh to show it. The Frisch-Waugh Theorem (whose proof is quite straightforward but out of the scope of the course, so you can easily check in any textbook or even online) establishes that if we are interested in β_1 in the following regression:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + U_i, \tag{17}$$

we can apply two different procedures that provide exactly (numerically) the same result. The first one is OLS on the whole regression. The second is to regress X_{1i}

and Y_i on X_{2i}, obtain the residuals of the two regressions, namely V_i and E_i respectively, and then estimate the following regression:

$$E_i = \beta_1 V_i + U_i. \tag{18}$$

Using this result, and noting that, since the regression of W_i on the treatment variable D_i provides a zero coefficient given independence (through randomization), the resulting regression would be:

$$E_i = Y_i - \gamma_Y W_i = \tilde{\beta}_0 + \beta D_i + \tilde{U}_i, \tag{19}$$

where γ_Y is the regression coefficient of Y_i on W_i, and $\tilde{\beta}_0 \equiv \mathbb{E}[Y_{0i} - \gamma_Y W_i]$. Since $\text{Var}(E_i) \leq \text{Var}(Y_i)$, then $\text{Var}(\tilde{U}_i) \leq \text{Var}(U_i)$.

A different role for additional regressors in randomized experiments is related to checking for randomization. In many real world examples of randomized experiments, there are serious questions about how well the randomization was implemented in practice. In such situation, adding some controls can provide additional credibility to the results. A previous step to that in practice is to check whether the possible additional controls have different averages in treated and control groups. A fast way of doing it is to compute the treatment effect on these variables and test whether we can reject that it equals zero. If it appears that treatment and control samples differ in a particular dimension, including this variable as a control could eliminate the resulting omitted variable bias.

Similarly, sometimes randomization is implemented conditional on observables. Regressors can be useful at the design stage (e.g. at the village level). This ensures that control units are not "contaminated" by treatment of treated units. In these cases, we need to further control for the variables used in the randomization design. This and the previous cases lead to the conditional independence situation, discussed in the next chapter.

C. Partial or Imperfect Compliance and Intention-to-Treat Analysis

So far, we have assumed that those in the treatment group all get the treatment and those in the control group do not. There are a number of reasons why things are often not as clean as this in practice. Those in the treatment group often cannot be forced to take the de-worming drugs or attend their training program or to take the offered savings package. Similarly, some in the control group may manage to get treatment because they complain or because close substitutes to the treatment are available outside the experiment. In the presence of imperfect

compliance, the probability of receiving treatment among the treatment group is less than one and/or it is more than zero for the control group.

For example, Kling, Liebman, and Katz (2007) provide an evaluation of the Moving To Opportunity (MTO) program in five US cities. This program gave some residents of public housing projects in disadvantaged neighborhoods the opportunity to move out of their public housing. The control group got no new assistance but there were two treatment groups. The S-group received a housing voucher they could use in private rental housing and the E-group the same but with use restricted to areas with poverty rates below 10%. The program is an opportunity to do something, nobody is forced to use the voucher. In fact, only 60% of the S-group and 47% of the E-group did.

The economic interest in this program is the following. It is well-known that cities tend to have residential sorting in which people with similar socioeconomic backgrounds live together. If there are externalities between neighbors then economic theory suggests this sorting may be inefficient. For example, are kids affected by growing up in a bad neighborhood or is their future affected solely by their household characteristics (which tend to be bad in a bad neighborhood)? With non-experimental evidence it has proved very hard to get credible evidence on this issue, but the experimental nature of the MTO program offers a chance to improve our knowledge on this important question.

Let D_i denote actual receipt of the treatment (using the voucher) and let Z_i denote being assigned to the treatment (receiving the voucher). So far we assumed that $Z_i = 1$ implied $D_i = 1$, and $Z_i = 0$ implied $D_i = 0$, but now we depart from this assumption. Individuals with $Z_i = 1$ but D_i are sometimes referred to as *no-shows*, because they did not show up to get the treatment, and individuals with $Z_i = 0$ but $D_i = 1$ are referred to as *cross-overs*.

The main concern here is that we are no-longer in a situation of independent treatment, as compliance can be endogenous to potential outcomes. Now:

$$Y_{1i}, Y_{0i} \not\perp\!\!\!\perp D_i, \tag{20}$$

but, in this case:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp Z_i. \tag{21}$$

The notation is not casual, as Z_i can be used as an instrumental variable, as discussed below. Alternatively, we can use Z_i as the treatment variable, instead

of D_i:

$$\alpha_{ITT} \equiv \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \tag{22}$$

This parameter is known as *intention-to-treat* effect.

D. Longer Run Interaction of Treatment and Intermediate Outcomes

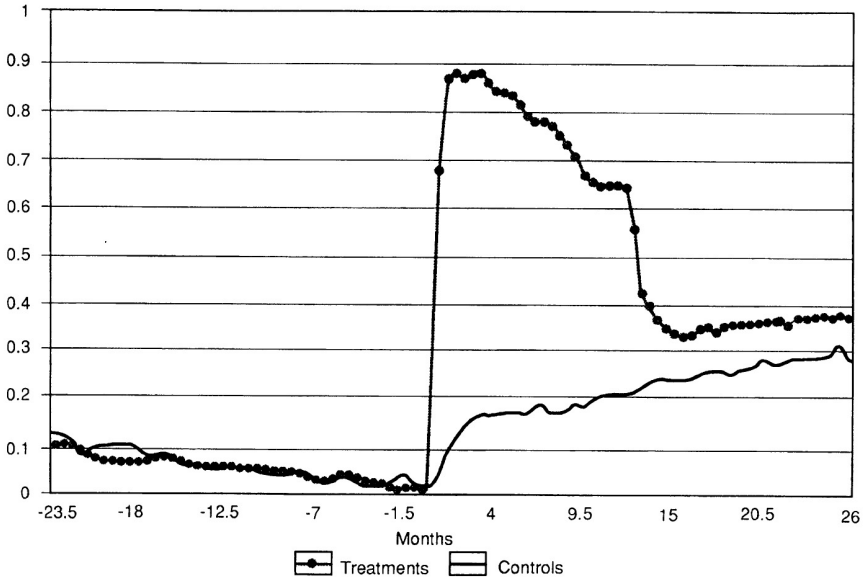
Ham and LaLonde (1996) analyze the National Supported Work program (NSW). The NSW was a training program designed in the U.S. in the mid 1970s to provide training and job opportunities to disadvantaged workers, as part of an experimental demonstration. Ham and LaLonde look at the effects of the NSW on women that volunteered for training. NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards. Eligibility requirements were to be unemployed, a long-term AFDC recipient, and have no preschool children. Participants were randomly assigned to treatment and control groups in 1976-1977. The experiment took place in 7 cities. Ham and LaLonde analyze data for 275 women in the treatment group and 266 controls. All volunteered in 1976.

Thanks to randomization, a simple comparison between employment rates of treatments and controls gives an unbiased estimate of the effect of the program on employment at different horizons. Figure 1 below, reproduced from Ham and LaLonde (1996) shows the effects. Initially, by construction there is a mechanical effect from the fact that treated women are offered a subsidized job. As apparent from the figure, compliance with the treatment is decreasing over time, as women can decide to drop from the subsidized job. The employment growth for controls is just a reflection of the program's eligibility criteria. Importantly, after the program ends, a 9 percentage points difference in employment rates is sustained in the medium run, at least until month 26 after the beginning of the program.

But Ham and LaLonde make an important additional point. Even though randomization allows researchers to evaluate the impact of the program on a particular outcome (employment) simply by comparing means, this is not true for any possible outcomes. In particular, if one is interested in the effect of the program on wages or on employment and unemployment durations, a comparison of means would provide a biased estimate of the effect of the program. This is because, as discussed above, the training program had an effect on employment rates of the treated.

To illustrate that, let W_i denote wages, let Y_i be an indicator variable that takes the value of one if the individual is employed, and zero if she is unemployed, and

FIGURE 1. EMPLOYMENT RATES OF AFDC WOMEN IN NSW DEMONSTRATION



Note: This figure corresponds to Figure 1 in Ham and LaLonde (1996)

let η_i denote the ability type, with $\eta_i = 1$ if the individual is skilled, and $\eta_i = 0$ if she is unskilled. Suppose that the treatment increases the employment rates of high skill and low skill workers, but the effect is of less intensity for the high skilled (as they were more likely to find a job anyway without the training program):

$$P(Y_i = 1|D_i = 1, \eta_i = 0) > P(Y_i = 1|D_i = 0, \eta_i = 0), \tag{23}$$

$$P(Y_i = 1|D_i = 1, \eta_i = 1) > P(Y_i = 1|D_i = 0, \eta_i = 1), \tag{24}$$

and:

$$\frac{P(Y_i = 1|D_i = 1, \eta_i = 0)}{P(Y_i = 1|D_i = 0, \eta_i = 0)} > \frac{P(Y_i = 1|D_i = 1, \eta_i = 1)}{P(Y_i = 1|D_i = 0, \eta_i = 1)}. \tag{25}$$

This implies that the frequency of low skill will be greater in the group of employed treatments than in the employed controls:

$$P(\eta_i = 0|Y_i = 1, D_i = 1) > P(\eta_i = 0|Y_i = 1, D_i = 0), \tag{26}$$

which is a way to say that η_i, which is unobserved, is not independent of D_i given $Y_i = 1$, although, unconditionally, $\eta_i \perp\!\!\!\perp D_i$. For this reason, a direct comparison of average wages between treatments and controls will tend to underestimate the effect of treatment on wages. In particular, consider the conditional effects:

$$\Delta_0 \equiv \mathbb{E}[W_i|Y_i = 1, D_i = 1, \eta_i = 0] - \mathbb{E}[W_i|Y_i = 1, D_i = 0, \eta_i = 0], \tag{27}$$

$$\Delta_1 \equiv \mathbb{E}[W_i|Y_i = 1, D_i = 1, \eta_i = 1] - \mathbb{E}[W_i|Y_i = 1, D_i = 0, \eta_i = 1]. \tag{28}$$

Our effect of interest is:

$$\Delta_{ATE} = \Delta_0 P(\eta_i = 0) + \Delta_1 P(\eta_i = 1), \tag{29}$$

whereas the comparison of average wages between treatments and controls gives:

$$\Delta_W = \mathbb{E}[W_i|Y_i = 1, D_i = 1] - \mathbb{E}[W_i|Y_i = 1, D_i = 0]. \tag{30}$$

In general, we shall have $\Delta_W < \Delta_{ATE}$. Indeed, it may not be possible to construct an experiment to measure the effect of training the unemployed on subsequent wages, i.e. it does not seem possible to experimentally undo the conditional correlation between D_i and η_i.

III. Matching

A. Selection Based on Observables and (Exact) Matching

There are many situations where experiments are too expensive, unfeasible, or unethical. A classical example is the analysis of the effects of smoking on mortality. Also, in experimental settings, often randomization is implemented conditional on observable characteristics. In any of these situations, we rely on observational data, which is unlikely to satisfy independence. In some situations, however, we can arguably defend the assumption of conditional independence, which is also referred to as selection based on observables:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|X_i. \tag{31}$$

When there is selection based on observables, the simple comparison of treatment and control averages does not deliver our treatment effects of interest, as the selection bias is not equal to zero. The problem is that the controls are not a counterfactual of treated in the absence of treatment, because the two groups differ in characteristics that are correlated with the outcome. As we discussed above, the average treatment effect is given by:

$$\alpha_{ATE} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i])dF(X_i), \tag{32}$$

and the average treatment effect on the treated is:

$$\alpha_{TT} = \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i])dF(X_i|D_i = 1). \tag{33}$$

What the above expressions do is to compare average outcomes for individuals with the same characteristics, and then integrate over the distribution of characteristics. In other words, for each treated (or control) unit, it imputes a counterfactual potential outcome when untreated (treated) obtained from individuals

in the control (treatment) group that share the same characteristics. This imputation is called (exact) *matching*, as it links each group of individuals in the treatment group with their counterparts in the control group (the "exact" qualifier is associated to the exercise of matching observations for every single value of X_i —below we review some alternatives that are more feasible when samples are not very large or the number of different combinations of covariate values is large or infinite). Following the discussion above, the reason why Equations (32) and (33) are unbiased representations of α_{ATE} and α_{TT} is that, since the selection is based on observables, for a given X_i the assignment to treatment and control groups is random, and $\mathbb{E}[Y_i|D_i = 1, X_i] = \mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_{1i}|X_i]$ and analogously $\mathbb{E}[Y_i|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|D_i = 1, X_i] = \mathbb{E}[Y_{0i}|X_i]$.

B. The Common Support Condition

An essential condition for matching is that there is some observation to match. In other words, for each possible value of X_i, there should be individuals in the treatment and control group for which we can average outcomes. This requirement is called the *common support condition*. Formally, this condition is stated as:

$$0 < P(D_i = 1|X_i) < 1 \quad \text{for all } X_i \text{ in its support.} \tag{34}$$

For example, assume that X_i is a single covariate. Denote the support of X_i by (X_{min}, X_{max}). Assume that the support for the subpopulation of treated subjects is $(X_{min}, \overline{X})$, and the support for the controls is $(\underline{X}, X_{max})$, with $\overline{X} > \underline{X}$. Then:

$$P(D_i = 1|X_i) = \begin{cases} 1 & \text{if } X_{min} \leq X < \underline{X} \\ p \in (0,1) & \text{if } \underline{X} \leq X \leq \overline{X} \\ 0 & \text{if } \overline{X} < X \leq X_{max} \end{cases}. \tag{35}$$

Given these assumptions, $\mathbb{E}[Y_i|D_i = 1, X_i]$ is only identified for values of X_i in the range $(X_{min}, \overline{X})$, and $\mathbb{E}[Y_i|D_i = 0, X_i]$ is only identified for values of X_i in the range $(\underline{X}, X_{max})$. Thus, we can only compute the difference $\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$ for values of X_i in the intersection range $(\underline{X}, \overline{X})$, which implies that α_{ATE} and α_{TT} are not identified.

C. Propensity Score Matching

Sometimes, the set of variables on which we need to do the matching is too large or multivariate. However, not all information included in X_i is relevant to obtain independence. Rosenbaum and Rubin (1983) introduced the *propensity score matching*, which is a method for reducing dimensionality based in the information

that is relevant for independence. They define the propensity score, $\pi(X_i)$, as:

$$\pi(X_i) \equiv P(D_i = 1|X_i). \tag{36}$$

Then, they note that $\pi(X_i)$ is a sufficient statistic for the distribution of D_i by construction, as:

$$\begin{aligned} P(D_i = 1|Y_{1i}, Y_{0i}, \pi(X_i)) &= \mathbb{E}[D_i|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\mathbb{E}[D_i|Y_{1i}, Y_{0i}, X_i]|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\mathbb{E}[D_i|X_i]|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[P(D_i = 1|X_i)|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \mathbb{E}[\pi(X_i)|Y_{1i}, Y_{0i}, \pi(X_i)] \\ &= \pi(X_i), \end{aligned} \tag{37}$$

where the third equality is obtained by using the conditional independence assumption. As a result, conditional independence given X_i is equivalent to conditional independence given the propensity score $\pi(X_i)$:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|X_i \quad \Leftrightarrow \quad Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i|\pi(X_i). \tag{38}$$

Thus, instead of matching exactly (based on the different values of X_i), we can match all observations with the same propensity score, whether or not they share the same covariates X_i. That is the propensity score matching.

Intuitively, we can bunch all X_i that share the same propensity score together because then treated and control groups are not going to overrepresented in any of these characteristics. For example, consider the analysis of a training program offered to disadvantaged unemployed workers. Let X_i be a vector of race and gender. Let the propensity scores and distribution of characteristics be given by:

	Propensity score		Probability mass function	
	black	white	black	white
male	0.3	0.1	0.1	0.4
female	0.8	0.3	0.1	0.4

In this example, black male and white female have the same probability of receiving treatment. The fraction of treated black male in the population is $0.3 \times 0.1 = 0.03$ and the one of treated white female is $0.3 \times 0.4 = 0.12$. For controls, these fractions are 0.07 and 0.28. Thus, if we restrict the sample to black male and white female, we observe that $0.03/(0.03 + 0.12) = 20\%$ of treated individuals in

this subsample are black male, and 80% are white female, and the same in the control group, $0.07/(0.07+0.28) = 20\%$ and 80%. Thus, within this subsample, there is no selection bias, as the treated and control groups are representative of the same subpopulation (this is, also $0.1/(0.1+0.4) = 20\%$ of individuals in the subpopulation are black male, and 80% are white female).

This result suggests two-step procedures to estimate the treatment effects where first we estimate the propensity score, and then create the appropriate weighting. To do so, we rewrite α_{ATE} in terms of the propensity score. Under (unconditional) independence, we established in Chapter 1 that:

$$\beta = \alpha_{ATE} = \mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0] = \frac{\mathbb{E}[D_i Y_i]}{P(D_i=1)} - \frac{\mathbb{E}[(1-D_i)Y_i]}{P(D_i=0)}. \quad (39)$$

Thus, under conditional independence we can write:

$$\begin{aligned}
\mathbb{E}[Y_{1i} - Y_{0i}|X_i] &= \mathbb{E}[Y_i|D_i=1,X_i] - \mathbb{E}[Y_i|D_i=0,X_i] \\
&= \frac{\mathbb{E}[D_i Y_i|X_i]}{P(D_i=1|X_i)} - \frac{\mathbb{E}[(1-D_i)Y_i|X_i]}{P(D_i=0|X_i)} \\
&= \frac{\mathbb{E}[D_i Y_i|X_i]}{\pi(X_i)} - \frac{\mathbb{E}[(1-D_i)Y_i|X_i]}{1-\pi(X_i)} \\
&= \mathbb{E}\left[\frac{D_i Y_i}{\pi(X_i)} - \frac{(1-D_i)Y_i}{1-\pi(X_i)}\middle| X_i\right] \\
&= \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)(1-\pi(X_i))}\middle| X_i\right], \quad (40)
\end{aligned}$$

and:

$$\alpha_{ATE} = \mathbb{E}\left[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]\right] = \mathbb{E}\left[Y_i \frac{D_i - \pi(X_i)}{\pi(X_i)[1-\pi(X_i)]}\right]. \quad (41)$$

This expression constitute an estimand to make inference on by one of the estimation methods described below.

To gain intuition on this expression, note that observations with $D_i = 1$ have a contribution of $Y_i/\pi(X_i)$, whereas observations with $D_i = 0$ have a contribution of $-Y_i/(1-\pi(X_i))$. In practice, what we are computing is a weighted average difference between observations in the treated group and in the control group. In our example before, we relatively underweight observations of black female in the treated group (their weight is $1/0.8 = 1.25$) because they are overrepresented (the overall fraction of treated is $0.1\times0.4+0.3\times(0.1+0.4)+0.8\times0.1 = 0.27$, and the fraction of them that are black female is $(0.8\times0.1)/0.27 = 29.6\%$, much larger than the 10% they represent overall), whereas white male are overweighted in this group (their weight is $1/0.1 = 10$), as they are underrepresented ($(0.1\times0.4)/0.27 = 14.8\% < 40\%$). The reverse is true for the control group.

D. Estimation Methods

The first and simplest method for matching estimation only works if X_i is discrete and relatively low-dimensional. Suppose X_i is indeed discrete and takes on J possible values $\{x_j\}_{j=1}^J$, and we have a sample of N observations $\{X_i\}_{i=1}^N$. Let N^j be the number of observations in cell j, N_ℓ^j be the number of observations in cell j with $D_i = \ell$, and $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D_i = \ell$. With this notation, $\bar{Y}_1^j - \bar{Y}_0^j$ is the sample counterpart of $\mathbb{E}[Y_i|D_i = 1, X_i = x_j] - \mathbb{E}[Y_i|D_i = 0, X_i = x_j]$, which can be used to obtain the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N^j}{N} \tag{42}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N_1^j}{N_1}. \tag{43}$$

Note that the formula for $\widehat{\alpha}_{TT}$ can also be written in the form:

$$\widehat{\alpha}_{TT} = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i - \bar{Y}_0^{j(i)}\right), \tag{44}$$

where $j(i)$ indicates the cell of X_i. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of untreated units in the same cell. In practice, this is a way of imputing the missing potential outcome for the treated individuals, and compute the average treatment effect for them. Note that this expression is the sample analog of Equation (33). We can proceed analogously with the propensity score $\pi(X_i)$ instead of the regressors X_i.

Alternatively, a straightforward way to perform propensity score matching estimation was proposed by Hirano, Imbens, and Ridder (2003). This method essentially estimates a sample analog of Equation (41), which we implement in two stages. In a first stage, we estimate $\widehat{\pi}(X_i)$ either non-parametrically or by means of a flexible parametric model like a Logit or Probit with polynomials, interactions, and the alike. In a second stage, we estimate the following quantity:

$$\widehat{\alpha}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \left(\frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1 - \widehat{\pi}(X_i)]} \right). \tag{45}$$

More generally, a matching estimator can be regarded as a way of constructing imputations for missing potential outcomes in a similar way, so that gains $Y_{1i} - Y_{0i}$

can be estimated for each unit. For example, in Equation (44), the imputation is:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k:D_k=0} Y_k \frac{\mathbb{1}\{X_k = X_i\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{X_\ell = X_i\}}. \tag{46}$$

More generally we compute:

$$\widehat{Y}_{0i} = \sum_{k:D_k=0} w(i,k) Y_k, \tag{47}$$

where different weighting schemes $w(i,k)$ determine different estimators.

The *nearest neighbor matching* uses the following weighting function:

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\}, \tag{48}$$

which, in words, means picking the individual k in the control group with the closest observables to the individual i in the treated group. Alternatively, the *radius matching* uses:

$$w(i,k) = \frac{\mathbb{1}\{||X_k - X_i|| < \varepsilon\}}{\sum_{\ell:D_\ell=0} \mathbb{1}\{||X_\ell - X_i|| < \varepsilon\}}, \tag{49}$$

for some threshold ε. In words, this procedure averages the observations from the control group with covariates within a window centered at X_i. And finally, the *kernel matching* uses:

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell:D_\ell=0} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)}, \tag{50}$$

where $\kappa(\cdot)$ is a kernel function that downweights distant observations, and γ_{N_0} is a bandwidth parameter. These procedures are generally implemented with replacement, meaning that each individual in the control group can be selected as a counterfactual for more than one individual in the treated group. Also they are typically applied to compute α_{TT}, but they are also applicable to α_{ATE}. And, furthermore, they can also be implemented on the propensity score $\pi(X_i)$ rather than the covariates X_i.

IV. Instrumental Variables

A. *Identification of Causal Effects in IV Settings*

Suppose that $(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i | X_i$, but we have an exogenous source of variation in D_i so that $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i | X_i$ that satisfies the relevance condition $Z_i \not\perp\!\!\!\perp D_i | X_i$.

In that situation, we can use the variation in Z_i to identify α_{ATE} under certain circumstances. Selection on observables can be regarded as a special case in which $Z_i = D_i$. For simplicity, we do most of the analysis below considering a single binary instrument $Z_i \in \{0, 1\}$, and we abstract from including other covariates. Here, it is crucial to distinguish the two cases we have been discussing so far: homogeneous and heterogeneous treatment effects.

B. Homogeneous Treatment Effects

Recall that in the homogeneous treatment effects world, the treatment effect is the same for all individuals, $Y_{1i} - Y_{0i} = \beta = \alpha_{ATE} = \alpha_{TT}$ for all individuals. In this case, the availability of an instrumental variable allows us to identify α_{ATE}. This is the traditional situation in econometric models with endogenous explanatory variables (IV regression). In particular, let $Y_i = \beta_0 + \beta D_i + U_i$ as in previous chapters. In this context, $D_i \not\perp\!\!\!\perp U_i$ given that $D_i \not\perp\!\!\!\perp Y_{0i}$. However, we use Z_i as an instrument for D_i in a just-identified fashion. Thus, the IV coefficient is given by:

$$\alpha = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, D_i)}. \tag{51}$$

Operating the numerator as in previous chapters we obtain:

$$\begin{aligned}
\text{Cov}(Z_i, Y_i) &= \mathbb{E}[Y_i Z_i] - \mathbb{E}[Y_i]\,\mathbb{E}[Z_i] \\
&= \mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1) \\
&\quad - \{\mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1) + \mathbb{E}[Y_i|Z_i = 0](1 - P(Z_i = 1))\}P(Z_i = 1) \\
&= \{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]\}P(Z_i = 1)(1 - P(Z_i = 1)).
\end{aligned} \tag{52}$$

Likewise, the denominator is:

$$\text{Cov}(Z_i, D_i) = \{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]\}P(Z_i = 1)(1 - P(Z_i = 1)). \tag{53}$$

Thus, the IV coefficient is:

$$\alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}, \tag{54}$$

which is known as the *Wald estimand*. This estimand can also be derived by noting that:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = Y_{0i} + \alpha D_i. \tag{55}$$

Since $Y_{0i} \perp\!\!\!\perp Z_i$, then:

$$\left.\begin{aligned}
\mathbb{E}[Y_i|Z_i = 1] &= \mathbb{E}[Y_{0i}] + \alpha\,\mathbb{E}[D_i|Z_i = 1] \\
\mathbb{E}[Y_i|Z_i = 0] &= \mathbb{E}[Y_{0i}] + \alpha\,\mathbb{E}[D_i|Z_i = 0]
\end{aligned}\right\} \Rightarrow \alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}. \tag{56}$$

Identification obviously requires that $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] \neq 0$, which is the relevance condition. All in all, we are obtaining the effect of D_i on Y_i through the effect of Z_i because Z_i only affects Y_i through D_i (exclusion restriction).

C. Heterogeneous Treatment Effects

In the heterogeneous case, the availability of instrumental variables is not sufficient to identify a causal effect (e.g. α_{ATE}). An additional assumption that helps in the identification of causal effects is the following *monotonicity condition*: any person that is willing to treat if assigned to the control group is also willing to treat if assigned to the treatment group. The plausibility of this assumption depends on the context of the application. Under monotonicity, the IV coefficient coincides with the average treatment effect for those individuals whose value of D_i would change when changing the value of Z_i, which is known as the *local average treatment effect* (LATE).

The monotonicity condition is well illustrated implementing the potential outcome notation also for the treatment variable. Let D_{0i} denote D_i when $Z_i = 0$, and let D_{1i} denote D_i when $Z_i = 1$ so that $D_{1i}, D_{0i} \perp\!\!\!\perp Z_i$. As we only observe $D_{\ell i}$, for individuals with $Z_i = \ell$, the combination of treatment and instrument define four *observable* groups. However, there are eight *potential* groups, depending on the value of the unobserved treatment status $D_{-\ell}$, which are listed in the following table:

Obs. type	Z	D	D_0	D_1	**Latent type**
Type 1	0	0	0	0	Never-taker
				1	Complier
Type 2	0	1	1	0	Defier
				1	Always-taker
Type 3	1	0	0	0	Never-taker
			1		Defier
Type 4	1	1	0	1	Complier
			1		Always-taker

For example, assume we are interested in the effect of college attendance (treatment) on wages (outcome). Because individuals with higher ability may be more likely to go to college and, for any given educational level, more likely to earn higher wages, independence does not hold neither conditionally nor unconditionally (we do not observe ability). Hence, to be able to identify a causal effect of college attendance on wages, we need an instrument. We consider proximity to a

college as an exogenous source of variation: it is associated with the cost of education, but plausibly uncorrelated with later outcomes. To make it dichotomous, we distinguish between being *far* and *close* from school. A complier is an individual that lives close to school and attends, but would not attend if she lived far, or one that does not attend school because she leaves far, but would have attended had she lived close. An individual that goes to school whether she lives close or far is an always-taker, and one that does not go to school whether she lives close or far is a never-taker. Defiers are those individuals that go to school being far, but would not go had they been close, or those who do not go being close, but would have gone had they been far. Monotonicity implies that there are no defiers.

To see that the availability of an instrumental variable is not enough to identify causal effects, consider the second derivation of the treatment effect for the homogeneous effects descried in Equation (56). Now we have:

$$\begin{aligned} \mathbb{E}[Y_i|Z_i=1] &= \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i}-Y_{0i})D_{1i}] \\ \mathbb{E}[Y_i|Z_i=0] &= \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i}-Y_{0i})D_{0i}], \end{aligned} \tag{57}$$

which implies:

$$\begin{aligned} \mathbb{E}[Y_i|Z_i=1] - \mathbb{E}[Y_i|Z_i=0] &= \mathbb{E}[(Y_{0i}-Y_{0i})(D_{1i}-D_{0i})] \\ &= \mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1]P(D_{1i}-D_{0i}=1) \\ &\quad - \mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=-1]P(D_{1i}-D_{0i}=-1). \end{aligned} \tag{58}$$

In this expression, $\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]$ could be negative even if the causal effect is positive for all units, as long as the fraction of defiers, $P(D_{1i}-D_{0i}=-1)$, is sufficiently large. Assuming monotonicity, we avoid this problem: in this case, the second term is zero, and we define $\mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1]$ as the local average treatment effect (LATE), which is as much as we can identify as we discuss below.

D. Imperfect Compliance and IV

One possibility is to make an even stronger assumption than monotonicity. In particular, assume an *eligibility rule* of the form:

$$P(D_i=1|Z_i=0)=0. \tag{59}$$

This rule implies that individuals with $Z_i=0$ are denied treatment (observable types 2, 3B, and 4B are ruled out). This situation occurs in the most standard form of imperfect compliance, discussed in Chapter 2: some individuals are assigned to treatment, but they endogenously decide whether to take it or not.

Under this eligibility rule:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i=1] &= \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i}-Y_{0i})D_i|Z_i=1] \\ &= \mathbb{E}[Y_{0i}] + \mathbb{E}[Y_{1i}-Y_{0i}|D_i=1,Z_i=1]P(D_i=1|Z_i=1) \\ &= \mathbb{E}[Y_{0i}] + \mathbb{E}[Y_{1i}-Y_{0i}|D_i=1]P(D_i=1|Z_i=1), \end{aligned} \tag{60}$$

where the last equality holds because $D_i = 1$ is a sufficient statistic to indicate that $Z_i = 1$ since $P(D_i = 1|Z_i = 0) = 0$. Likewise:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i=0] &= \mathbb{E}[Y_{0i}] + \mathbb{E}[(Y_{1i}-Y_{0i})D_i|Z_i=0] \\ &= \mathbb{E}[Y_{0i}] + \mathbb{E}[Y_{1i}-Y_{0i}|D_i=1,Z_i=0]P(D_i=1|Z_i=0) \\ &= \mathbb{E}[Y_{0i}]. \end{aligned} \tag{61}$$

Thus, we can identify the average treatment effect on the treated in this case, as:

$$\begin{aligned}\alpha_{TT} &= \mathbb{E}[Y_{1i}-Y_{0i}|D_i=1] \\ &= \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{P(D_i=1|Z_i=1)} \\ &= \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]}, \end{aligned} \tag{62}$$

which is the Wald estimand, as $\mathbb{E}[D_i|Z_i = 0] = 0$ by the assumption in (59).

E. Local Average Treatment Effects (LATE)

As discussed above, under monotonicity, Equation (58) reduces to:

$$\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0] = \mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1]P(D_{1i}-D_{0i}=1). \tag{63}$$

Also, assuming $D_{1i}, D_{0i} \perp\!\!\!\perp Z_i$ (implying the proportions of compliers, always-takers, and never-takers in the subsample with $Z_i = 1$ coincides with the one in the subsample with $Z_i = 0$) along with monotonicity, we have:

$$\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0] = \mathbb{E}[D_{1i}-D_{0i}] = P(D_{1i}-D_{0i}=1). \tag{64}$$

Thus, the causal effect that we can identify is:

$$\alpha_{LATE} \equiv \mathbb{E}[Y_{1i}-Y_{0i}|D_{1i}-D_{0i}=1] = \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}, \tag{65}$$

which is given again by the Wald estimand. Imbens and Angrist (1994) called this parameter a local average treatment effect, because averages treatment effects on the subsample of compliers. Importantly, different instrumental variables

lead to different parameters, even under instrument validity, which is counter to standard GMM thinking. This concept changed radically the way we think of and understand IV. As noted, the identified coefficient is the average treatment effect for compliers. Thus, when selecting an instrument, on top of thinking about relevance and orthogonality conditions, the researcher needs to think about the potential group of compliers selected by the instrument.

The most relevant LATEs are those based on instruments that are policy variables. For example, in the college attendance example before, the identified LATE (the effect of schooling for those individuals changing their enrollment based on distance to college) is very relevant for a subsidy policy, even if it is not a good measurement of the average return to education in the whole population.

As a final remark, what happens if there are no compliers? In the absence of defiers, the probability of compliers satisfies $P(D_{1i} - D_{0i} = 1) = \mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$, so the lack of compliers implies lack of instrument relevance, and, hence, underidentification. This is natural, because if the population is formed of never-takers and always-takers, there is no role to be played by the instrument.

F. Conditional Estimation with Instrumental Variables

So far we abstracted from the fact that the validity of the instrument may only be conditional on X_i: it may be that $(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp Z_i$, but the following does:

$$\begin{array}{ll} (Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i|X_i & \text{(conditional independence)} \\ Z_i \not\perp\!\!\!\perp D_i|X_i & \text{(conditional relevance)} \ . \end{array} \tag{66}$$

For example, in the analysis of the returns to college, Z_i is an indicator of proximity to college. The problem is that Z_i is not randomly assigned but chosen by parents, and this choice may depend on characteristics that subsequently affect wages. The validity of Z_i may be more credible if we can condition on family background, X_i.

In the linear version of the problem we can estimate using a two-stage procedure: first regress D_i on Z_i and X_i, so that we get $\widehat{D}_i$, and in the second stage we regress Y_i on $\widehat{D}_i$ and X_i. In general, we now have a conditional LATE given X_i:

$$\gamma(X_i) \equiv \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} - D_{0i} = 1, X_i], \tag{67}$$

and a conditional IV estimator:

$$\beta(X_i) \equiv \frac{\mathbb{E}[Y_i|Z_i = 1, X_i] - \mathbb{E}[Y_i|Z_i = 0, X_i]}{\mathbb{E}[D_i|Z_i = 1, X_i] - \mathbb{E}[D_i|Z_i = 0, X_i]}. \tag{68}$$

To get an aggregate effect, we proceed differently depending on whether the effects

are homogeneous or heterogeneous. In the homogeneous case:

$$Y_{1i} - Y_{0i} = \beta(X_i) \quad \forall i. \tag{69}$$

In the heterogeneous case, it makes sense to consider an average treatment effect for the overall subpopulation of compliers:

$$\begin{aligned}\beta_C &\equiv \int \beta(X_i) \frac{P(compliers|X_i)}{P(compliers)} dF(X_i) \\ &= \int \{\mathbb{E}[Y_i|Z_i = 1, X_i] - \mathbb{E}[Y_i|Z_i = 0, X_i]\} \frac{1}{P(compliers)} dF(X_i),\end{aligned} \tag{70}$$

where:

$$P(compliers) = \int \{\mathbb{E}[D_i|Z_i = 1, X_i] - \mathbb{E}[D_i|Z_i = 0, X_i]\} \, dF(X_i). \tag{71}$$

Intuitively, in the top row we use the Bayes' Theorem to rewrite the density of X_i conditional on being a complier. Replacing (71) into (70) yields:

$$\beta_C = \frac{\int \{\mathbb{E}[Y_i|Z_i = 1, X_i] - \mathbb{E}[Y_i|Z_i = 0, X_i]\} \, dF(X_i)}{\int \{\mathbb{E}[D_i|Z_i = 1, X_i] - \mathbb{E}[D_i|Z_i = 0, X_i]\} \, dF(X_i)}, \tag{72}$$

which can be estimated as a ratio of matching estimators (Frölich, 2003).

V. Regression Discontinuity

A. *The fundamental RD assumption*

In what we have seen so far, the main assumption in the matching context is conditional independence, $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|X_i$, whereas in the IV context we assume orthogonality and relevance of the instrument, $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i|X_i$ and $D_i \not\perp\!\!\!\perp Z_i|X_i$ respectively. The relevance condition can also be expressed as $P(D_i = 1|Z_i = z) \neq P(D_i = 1|Z_i = z')$ for some $z \neq z'$. In *regression discontinuity* (RD) we consider a situation where there is a continuous variable Z_i that is not necessarily a valid instrument (it does not satisfy the exogeneity assumption), but that it is such that treatment assignment is a discontinuous function of Z_i. The basic asymmetry on which identification rests is discontinuity in the dependence of D_i on Z_i but continuity in the dependence of (Y_{1i}, Y_{0i}) on Z_i. RD methods have much potential in economic applications because geographic boundaries or program rules (e.g. eligibility thresholds) often create usable discontinuities.

More formally, discontinuity in treatment assignment but continuity in potential

outcomes means that there is at least a known value $z = z_0$ such that:

$$\lim_{z \to z_0^+} P(D_i = 1|Z_i = z) \neq \lim_{z \to z_0^-} P(D_i = 1|Z_i = z) \tag{73}$$

$$\lim_{z \to z_0^+} P(Y_{ji} \leq r|Z_i = z) = \lim_{z \to z_0^-} P(Y_{ji} \leq r|Z_i = z) \quad (j = 0, 1) \tag{74}$$

Implicit regularity conditions are: (i) the existence of the limits, and (ii) that Z_i has positive density in a neighborhood of z_0. We abstract from conditioning covariates for the time being for simplicity.

Early RD literature in Psychology (e.g. Cook and Campbell, 1979) distinguishes between *sharp* and *fuzzy* designs. In the former, D_i is a deterministic function of Z_i:

$$D_i = \mathbb{1}\{Z_i \geq z_0\}, \tag{75}$$

whereas in the latter is not. The sharp design can be regarded as a special case of the fuzzy design, but one that has different implications for identification of treatment effects. In the sharp design:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] &= 1 \\ \lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z] &= 0. \end{aligned} \tag{76}$$

B. Homogeneous Treatment Effects

Like in the IV setting, the case of homogeneous treatment effects is useful to present the basic RD estimator. Suppose that $\alpha = Y_{1i} - Y_{0i}$ is constant, so that:

$$Y_i = \alpha D_i + Y_{0i} \tag{77}$$

Taking conditional expectations given $Z_i = z$ and left-side and right-side limits:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] &= \alpha \lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_{0i}|Z_i = z] \\ \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z] &= \alpha \lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_{0i}|Z_i = z], \end{aligned} \tag{78}$$

which leads to the consideration of the following RD parameter:

$$\alpha = \frac{\lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z]}{\lim_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z]}, \tag{79}$$

which is determined provided the relevance condition in Equation (73) is satisfied, and equals α provided the independence condition in Equation (74) holds.

In the case of a sharp design, the denominator is unity so that:

$$\alpha = \lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z], \tag{80}$$

which can be regarded as a matching-type situation, in the same way that the general case can be regarded as an IV-type situation. So the basic idea is to obtain a treatment effect by comparing the average outcome left of the discontinuity with the average outcome to the right of discontinuity, relative to the difference between the left and right propensity scores. Intuitively, considering units within a small interval around the cutoff point is similar to a randomized experiment at the cutoff point.

C. Heterogeneous Treatment Effects

Now suppose that:

$$Y_i = \alpha_i D_i + Y_{0i}. \tag{81}$$

It is useful again to distinguish sharp and fuzzy designs.

Sharp design. In the sharp design, since $D_i = \mathbb{1}\{Z_i \geq z_0\}$ we have:

$$\mathbb{E}[Y_i|Z_i = z] = \mathbb{E}[\alpha_i|Z_i = z]\, \mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_{0i}|Z_i = z]. \tag{82}$$

In other words, conditioning on a value z for Z_i, individuals are treated if $z \geq z_0$, and thus we observe $Y_i = Y_{1i} = \alpha_i + Y_{0i}$, and untreated if $z \leq z_0$, in which case we observe $Y_i = Y_{0i}$. Thus, to obtain an average treatment effect for individuals at the threshold value z_0, that is, α_{RD} defined as:

$$\alpha_{RD} \equiv \mathbb{E}[\alpha_i|Z_i = z_0], \tag{83}$$

we rewrite (82) as:

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i = z] &= \mathbb{E}[\alpha_i|Z_i = z]\, \mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_{0i}|Z_i = z] \pm \mathbb{E}[\alpha_i|Z_i = z_0]\, \mathbb{1}\{z \geq z_0\} \\
&= \alpha_{RD}\, \mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_{0i}|Z_i = z] \\
&\qquad + (\mathbb{E}[\alpha_i|Z_i = z] - \mathbb{E}[\alpha_i|Z_i = z_0])\, \mathbb{1}\{z \geq z_0\} \\
&\equiv \alpha_{RD} D_i + k_{z_0}(z).
\end{aligned} \tag{84}$$

This equation corresponds to a situation of selection on observables, and the term $k_{z_0}(z)$ "controls" for the selection bias (this type of functions are indeed known as a control functions, and including them in the regression is know as a *control*

function approach). Thus, the OLS population coefficient on D_i in the equation:

$$Y_i = \alpha_{RD} D_i + k_{z_0}(Z_i) + w_i \tag{85}$$

equals $\mathbb{E}[\alpha_i | Z_i = z_0]$, which is the causal effect of interest (an average treatment effect for individuals with Z_i right below or above the discontinuity).

The control function $k_{z_0}(z)$ is nonparametrically identified (e.g. including a high-order polynomial in Z_i —or $Z_i - z_0$— in the OLS regression interacted with a dummy $\mathbb{1}\{Z_i \geq z_0\}$). Note that if the treatment effect is homogeneous, $k(z)$ coincides with $\mathbb{E}[Y_{0i} | Z_i = z]$, but not in general.

Fuzzy design. In the fuzzy design, D_i not only depends on $\mathbb{1}\{Z_i \geq z_0\}$, but also on other unobserved variables. Thus, D_i is an endogenous variable in Equation (85). However, we can still use $\mathbb{1}\{Z_i \geq z_0\}$ as an instrument for D_i in such equation to identify α_{RD}, at least in the homogeneous case. The connection between the fuzzy design and the instrumental variables perspective was first made explicit in van der Klaaw (2002).

Next, we discuss the interpretation of α_{RD} in the fuzzy design with heterogeneous treatment effects, under two different assumptions. Consider first the weak conditional independence assumption:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | Z_i = z \quad \text{for } z \text{ near } z_0, \tag{86}$$

that is, for $z = z_0 \pm e$, where e is an arbitrarily small positive number, or simply:

$$F(Y_{ji} | D_i = 1, Z_i = z_0 \pm e) = F(Y_{ji} | Z_i = z_0 \pm e) \quad (j = 0, 1). \tag{87}$$

Thus, we are assuming that treatment assignment is exogenous in the neighborhood of z_0. An implication is:

$$\mathbb{E}[\alpha_i D_i | Z_i = z_0 \pm e] = \mathbb{E}[\alpha_i | Z_i = z_0 \pm e] \, \mathbb{E}[D_i | Z_i = z_0 \pm e]. \tag{88}$$

Proceeding as before, we have:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[Y_i | Z_i = z] &= \lim_{z \to z_0^+} \mathbb{E}[\alpha_i | Z_i = z] \, \mathbb{E}[D_i | Z_i = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_{0i} | Z_i = z] \\ \lim_{z \to z_0^-} \mathbb{E}[Y_i | Z_i = z] &= \lim_{z \to z_0^-} \mathbb{E}[\alpha_i | Z_i = z] \, \mathbb{E}[D_i | Z_i = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_{0i} | Z_i = z]. \end{aligned} \tag{89}$$

Noting that $\lim_{z \to z_0^+} \mathbb{E}[\alpha_i | Z_i = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha_i | Z_i = z] = \alpha_{RD}$, subtracting one

equation from the other, and rearranging the terms we obtain:

$$\begin{aligned}\alpha_{RD} &\equiv \mathbb{E}[Y_{1i} - Y_{0i}|Z_i = z_0] \\ &= \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D_i|Z_i = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D_i|Z_i = z]}.\end{aligned} \tag{90}$$

That is, the RD parameter can be interpreted as the average treatment effect at z_0.

Hahn, Todd, and van der Klaaw (2001) also consider an alternative LATE-type of assumption. Let D_{zi} be the potential assignment indicator associated with $Z_i = z$, and for some $\bar{\varepsilon} > 0$ and any pair $(z_0 - \varepsilon, z_0 + \varepsilon)$ with $0 < \varepsilon < \bar{\varepsilon}$ suppose the local monotonicity assumption:

$$D_{z_0+\varepsilon,i} \geq D_{z_0-\varepsilon,i} \text{ for all units } i \text{ in the population.} \tag{91}$$

Sometimes, the local conditional independence assumption could be problematic, especially in fuzzy designs, but the monotonicity assumption is not. In such case, it can be shown that α_{RD} identifies the local average treatment effect at $z = z_0$:

$$\alpha_{RD} = \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_1 - Y_0|D_{z_0+\varepsilon} - D_{z_0-\varepsilon} = 1] \tag{92}$$

that is, the ATE for the units for whom treatment changes discontinuously at z_0. If the policy is a small change in the threshold for program entry, the LATE parameter delivers the treatment effect for the subpopulation affected by the change, so that in that case it would be the parameter of policy interest.

D. Estimation Strategies

There are parametric and semiparametric estimation strategies. Hahn *et al.* (2001) suggested the following local estimator. Let $S_i \equiv \mathbb{1}\{z_0 - h < Z_i < z_0 + h\}$ where $h > 0$ denotes the bandwidth, and consider the subsample such that $S_i = 1$. The proposed estimator is the IV regression of Y_i on D_i using $W_i \equiv \mathbb{1}\{z_0 < Z_i < z_0 + h\}$ as an instrument, applied to the subsample with $S_i = 1$:

$$\widehat{\alpha}_{RD} = \frac{\widehat{\mathbb{E}}[Y_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[Y_i|W_i = 0, S_i = 1]}{\widehat{\mathbb{E}}[D_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[D_i|W_i = 0, S_i = 1]}. \tag{93}$$

In sharp designs, the denominator equals 1. This estimator has a poor boundary performance. An alternative is based on Equation (85). In the case of a sharp design, OLS provides consistent estimates of α_{RD}, but in the fuzzy design D_i is endogenous. In that context, we would typically use $\mathbb{1}\{Z_i \geq z_0\}$ as an instrument for D_i. These regression methods, not local to data points near the threshold, are implicitly predicated on the assumption of homogeneous treatment effects.

E. Conditioning on Covariates

Even if the RD assumption is satisfied unconditionally, conditioning on covariates may mitigate the heterogeneity in treatment effects, hence contributing to the relevance of RD estimated parameters, which otherwise are "very local". Covariates may also make the local conditional exogeneity assumption more credible.

VI. Difference in Differences

A. The Setup

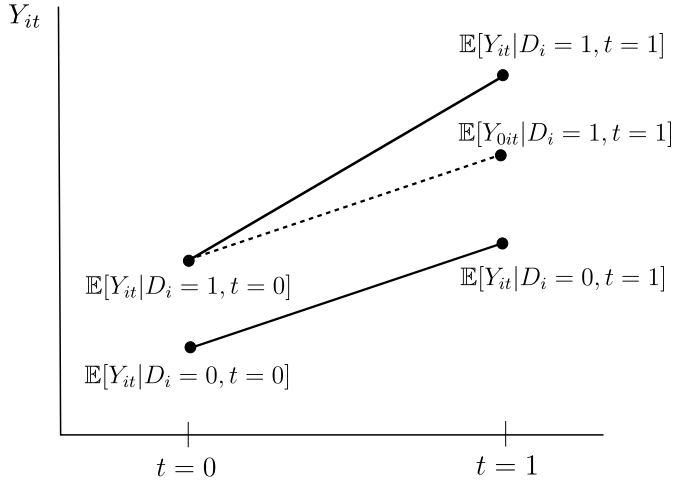
With data from a randomized experiment, the simple comparison of the mean outcome in treatment and control groups (which we can define here as the "difference" estimator) provides an unbiased and consistent estimate of the average treatment effect, as discussed in Chapter 2. This is so because the randomization ensures there are no systematic differences in any "pre-treatment" variables, and, hence, confounding factors are balanced.

In subsequent chapters we have dealt with deviations from the independence assumption. In Chapter 3, and in sharp RD designs in Chapter 5, we proposed different techniques that balance out systematic differences among treated and control units, creating comparable groups, and, thus, ruling out confounders. In Chapter 4 and fuzzy RD designs in Chapter 5 we tried to get causal effects by using instrumental variables. However, good instruments are hard to find, and we would like to have other techniques to rule out unobserved confounders.

The approach in this chapter, which builds on the toolkit developed in Chapter 6, follows an approach that is closer to the first of the two broad approaches described in the previous paragraph. Linking Chapter 6 to treatment effects approaches, we propose an alternative method to eliminate confounders that are fixed over time (like a fixed effect), using repeated observations over time. We assume that, even though treated and control groups are not comparable, the evolution of the outcome pre- and post-treatment would be the same in the absence of treatment. In other words, we assume that treated and control groups have the same counterfactual *trends*, even if the levels differ. In this case, we use data on treatment and control groups before the treatment to estimate the pre-treatment difference between these groups and then compare this difference with the difference in average outcomes after the treatment group received the treatment. Intuitively, the two differences would be equal in the absence of treatment, and the extra difference is imputed to the treatment effect.

The figure below illustrates this discussion. Let Y_{it} denote the observed outcome

for individual i in period $t \in \{0, 1\}$, and let $D_i = 1$ if the individual is in the treated group, with $D_i = 0$ otherwise. Note that we did not subscript D_i by time in this notation, as $D_{it} = 0$ when $t = 0$ for both treated and untreated individuals. For treated individuals we observe $\mathbb{E}[Y_{it}|D_i = 1, t = 0] = \mathbb{E}[Y_{0it}|D_i = 1, t = 0]$, because at $t = 0$ no observation is treated, and $\mathbb{E}[Y_{it}|D_i = 1, t = 1] = \mathbb{E}[Y_{1it}|D_i = 1, t = 1]$, because these individuals are treated at $t = 1$. Likewise, for controls, we observe $\mathbb{E}[Y_{it}|D_i = 0, t = 0] = \mathbb{E}[Y_{0it}|D_i = 0, t = 0]$ as well, but, in this case, the mean observed in the second period is $\mathbb{E}[Y_{it}|D_i = 0, t = 1] = \mathbb{E}[Y_{0it}|D_i = 0, t = 1]$. What we do not observe is $\mathbb{E}[Y_{0it}|D_i = 1, t = 1]$, which we need to compute the average treatment effect on the treated:



What the figure suggests is to use the same trend observed for untreated individuals to predict the counterfactual trend for treated individuals in the absence of treatment. Thus, our prediction of the counterfactual value $\mathbb{E}[Y_{0it}|D_i = 1, t = 1]$ is:

$$\mathbb{E}[Y_{0it}|D_i = 1, t = 1] = \underbrace{\mathbb{E}[Y_{it}|D_i = 0, t = 1]}_{\text{level for controls at } t=1} + \underbrace{\{\mathbb{E}[Y_{it}|D_i = 1, t = 0] - \mathbb{E}[Y_{it}|D_i = 0, t = 0]\}}_{\text{difference in levels at } t=0 \text{ difference}}, \quad (94)$$

which builds on the fundamental assumption that $\mathbb{E}[Y_{0i1} - Y_{0i0}|D_i = 1] = \mathbb{E}[Y_{0i1} - Y_{0i0}|D_i = 0]$. This assumption is known as *the common trend assumption*, and, where there are multiple periods before treatment, it is typically checked by showing that trends before treatmend coincided. Hence, the difference in differences coefficient (which is an average treatment effect on the treated) is:

$$\begin{aligned}\beta &= \mathbb{E}[Y_{1it}|D_i = 1, t = 1] - \mathbb{E}[Y_{0it}|D_i = 1, t = 1] \\ &= \{\mathbb{E}[Y_{it}|D_i = 1, t = 1] - \mathbb{E}[Y_{it}|D_i = 1, t = 0]\} \\ &- \{\mathbb{E}[Y_{it}|D_i = 0, t = 1] - \mathbb{E}[Y_{it}|D_i = 0, t = 0]\}. \end{aligned} \quad (95)$$

Intuitively, β measures the difference between the increase in average observed outcomes for treated and the increase in average observed outcomes for controls.

B. Difference in Differences in the Regression Context

The difference in differences coefficient can be obtained as the β coefficient in the following regression:

$$Y_{it} = \beta_0 + \beta_D D_i + \beta_T T_{it} + \beta D_i T_{it} + U_{it}, \tag{96}$$

where $T_{it} = 1$ if individual i is treatment period $t = 1$, and $T_{it} = 0$ otherwise. With a proof that is very similar than those done in previous chapters, one can prove that β_0 is $\mathbb{E}[Y_{it}|D_i = 0, t = 0]$, $\beta_0 + \beta_D = \mathbb{E}[Y_{it}|D_i = 1, t = 0]$, $\beta_0 + \beta_T = \mathbb{E}[Y_{it}|D_i = 0, t = 1]$, and β is the difference in differences coefficient.

This regression model can be expanded in several ways. First, by including further periods, both before, and after the treatment. In such case, T_{it} is not a time dummy but, instead, a dummy that equals one in the post-treatment period. One could additionally include time effects, but the interaction term should be with the "post" dummy only. Second, the regression allows for controls, X_{it}. In this context, the difference between the regression coefficient and the difference in differences coefficient (obtained nonparametrically from differences in means) is analogous to the difference between matching and regression coefficients discussed in Chapter 3. Third, actually there is no need for panel data to estimate (96): repeated cross sections should suffice. However, in the repeated cross-section context, the researcher needs to sustain the assumption that the sample composition does not vary over time, which is satisfied by construction with panel data. Furthermore, panel data would allow to control for individual fixed effects in the same way we discussed in Chapter 6. Finally, some authors use the same regression setup to build *placebo exercises*. A placebo regression is a regression that simulates the difference in differences analysis but for a point in time or group of individuals that resemble the treatment period or group but that was actually not treated. It is a "placebo" in the sense that it looks as if treatment was administered, but it actually was not.

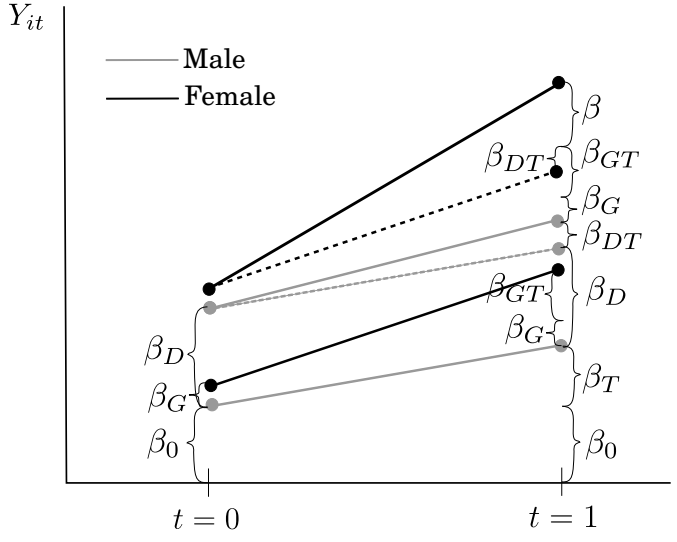
C. Triple Differences Model

Some authors pose *triple-differences* models, in which the difference in differences assumption does not hold, but the change in trends is assumed to be the same across sub-groups, some of which should be more affected than others. For example, let G_i denote the (say sociodemographic) group to which individual i

belongs. Then, the triple-differences model is:

$$Y_{it} = \beta_0 + \beta_D D_i + \beta_T T_{it} + \beta_G G_i + \beta_{GD} G_i D_i + \beta_{GT} G_i T_{it} + \beta_{DT} D_i T_{it} + \beta G_i D_i T_{it} + U_{it}. \quad (97)$$

For example, consider the analysis of maternity leave policies on labor supply. These policies affect young women but do not affect old women. In this context, even though the labor supply of old women is systematically different than that of young women (level difference), this systematic difference persists before and after the policy, and, therefore, we can use old women as a control group in a differences in differences setting. Now imagine that, at the same time that the maternity leave policy is introduced, a tax reform occurs that particularly affects the labor supply of young workers relative to old workers. This additional policy would constitute a counfounder that would break the common trend assumption, because it affects the treated group only in the "post-reform" period, as the maternity leave policy change. However, we have a different group of people, males, that are equally affected by the tax reform, but not affected by the maternity leave policy. In this context, we can use a difference-in-difference estimation for male to "remove" the effect of taxes from the composite effect on female (taxes plus maternal leave policy). In this case, the key assumption is that taxes affect male of different ages in the same way that they affect female. The triple difference coefficients are easily interpreted in the following figure:



D. Synthetic Control Methods

Consider the case in which we have several periods before treatment is implemented, and, thus, we can check the common trends assumption. For example,

consider the case where one state implements a policy and other states do not. With enough data, we could define as the control the state that has the most similar pre-trend compared to the treated group (or alternatively, all non-treated states). However, often no state is the perfect counterfactual for another.

Synthetic control methods use longitudinal data to build the weighted average of non-treated units that best reproduces the characteristics of the treated unit over time prior to the treatment. Thus, we build an artificial control that has the best possible pre-trend possible, and then we compute the difference in differences estimate using such synthetic control group.

References

Abadie, Alberto and Javier Gardeazabal (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113-132.

Angrist, Joshua D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-336.

Angrist, Joshua D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service using Social Security Data on Military Applicants," *Econometrica*, 66, 249-288.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics, An Empiricists Companion*, Princeton University Press.

Angrist, Joshua D. and Victor Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533-775.

Card, David E. and Alan B. Krueger (1994), "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772-793.

Cook, Thomas D. and Donald T. Campbell (1979), *Quasi-Experimentation: Design & Analysis. Issues for Field Settings*, Chicago: Rand McNally College Publishing Company.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir (2009), "Conditional Cash Transfer and School Dropout Rates," *Journal of Human Resources*, 44, 827-857.

Dehejia, Rajeev H. and Sadek Wahba (2002), "Propensity Score Matching

Methods for Non-Experimental Causal Studies," *Review of Economics and Statistics*, 84, 151-161.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit," CEPR Discussion Paper No. 6059.

Frölich, Markus (2003), *Program Evaluation and Treatment Choice*, Berlin-Heidelberg: Springer-Verlag.

Ham, John C. and Robert J. LaLonde (1996) "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training," *Econometrica*, 64, 175-205.

Heckman, James J. and Edward Vytlacil (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73, 669-738.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161-1189.

Imbens, Guido W. and Joshua D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

Lee, David S., and Thomas Lemieux (2009), "Regression Discontinuity Designs in Economics," Journal of Economic Literature, 48, 281-355.

Moffitt, Robert A. (2003) Means-Tested Transfer Programs in the United States. Chicago: The University of Chicago Press.

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Snow, John (1855), *On the Mode of Communication of Cholera*, Churchill, London. Reprinted by Hafner, New York, 1965.

Vytlacil, Edward (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331-341.