Introduction

This book on advanced econometrics is intended to familiarise the reader with technical developments in the area of econometrics known as *treatment effect estimation*, or *impact* or *policy evaluation*. In this book we try to combine intuitive reasoning in identification and estimation with econometric and statistical rigour. This holds especially for the complete list of stochastic assumptions and their implications in practice. Moreover, for both identification and estimation, we focus mostly on non-parametric methods (i.e. our methods are not based on specific pre-specified models or functional forms) in order to provide approaches that are quite generally valid. Graphs and a number of examples of evaluation studies are applied to explain how sources of exogenous variation can be explored when disentangling causality from correlation.

What makes the analysis of treatment effects different from more conventional econometric analysis methods, such as those covered, for example, in the textbooks of Cameron and Trivedi (2005), Greene (1997) or Wooldridge (2002)? A first major difference is that the three steps – definition of parameter of interest, identification and statistical modelling – are clearly separated. This helps first to define the objects one is interested in, and to clearly articulate the definition and interpretation of counterfactual outcomes. A second major difference is the focus on non-parametric identification and estimation. Even though parametric models might eventually be used in the empirical analysis, discussing identification without the need to impose – usually arbitrary – functional forms helps us to understand where the identifying power comes from. This permits us to link the identification strategy very tightly to the particular policy evaluation problem. A third, and also quite important, difference is the acknowledgement of possible treatment effect heterogeneity. Even though it would be interesting to model this heterogeneity of treatment effects, according to the standard literature we take it as being of unknown form: some individuals may benefit greatly from a certain intervention whereas some may benefit less, while others may even be harmed. Although treatment effects are most likely heterogeneous, we typically do not know the form of this heterogeneity. Nonetheless, the practitioner should always be aware of this heterogeneity, whereas (semi-)parametric regression models either do not permit it or do not articulate it clearly. For example, most of the instrumental variable (IV) literature simply ignores the problem of heterogeneity, and often people are not aware of the consequences of particular model or IV choices in their data analysis. This can easily render the presented interpretation invalid.

The book is oriented towards the main strands of recent developments, and it emphasises the reading of original articles by leading scholars. It does not and cannot substitute

for the reading of original articles, but it seeks to summarise most of the central aspects, harmonising notation and (hopefully) providing a coherent road map. Unlike some handbooks on impact evaluation, this book aims to impart a deeper understanding of the underlying ideas, assumptions and methods. This includes such questions as: what are the necessary conditions for the identification and application of the particular methods?; what is the estimator doing to the data?; what are the statistical properties, asymptotically and in finite samples, advantages and pitfalls, etc.? We believe that only a deeper understanding of all these issues (the economic theory that identifies the parameters of interest, the conditions of the chosen estimator or test and the behaviour of the statistical method) can finally lead to a correct inference and interpretation.

Quite comprehensive review articles, summarising a good part of the theoretical work that has been published in the last 15 years in the econometric literature,[1] include, for example, Imbens (2004), Heckman and Vytlacil (2007a), Heckman and Vytlacil (2007b), Abbring and Heckman (2007) and Imbens and Wooldridge (2009). See also Angrist and Pischke (2008). The classical area of application in economics was that of labour market research, where some of the oldest econometric reviews on this topic can be found; see Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999). Nowadays, the topic of treatment effect estimation and policy evaluation is especially popular in the field of poverty and development economics, as can be seen from the reviews of Duflo, Glennerster and Kremer (2008) and Ravallion (2008). Blundell and Dias (2009) try to reconcile these methods with the structural model approach that is standard in microeconometrics. Certainly, this approach has to be employed with care, as students could easily get the impression that treatment effect estimators are just semiparametric extensions of the well-known parameter estimation problems in structural models.

Before starting, we should add that this book considers the randomised control trials (RCT) only in the first chapter, and just as a general principle rather than in detail. The book by Guido W. Imbens and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, has appeared quite recently and deals with this topic in considerable detail. (See also the book Glennerster and Takavarasha (2013) on practical aspects of running RCTs). Instead, we have added to the chapters on the standard methods of matching, instrumental variable approach, regression discontinuity design and difference-in-differences more detailed discussions about the use of *propensity scores*, and we introduce in detail *quantile and distributional effects* and give an overview of the analysis of *dynamic treatment effects*, including sequential treatments and duration analysis. Furthermore, unlike the standard econometrics literature, we introduce (for the identification of causality structures) graph theory from the statistics literature, and give a (somewhat condensed) review of non-parametric estimation that is applied later on in the book.

[1] There exists an even more abundant statistical literature that we neither cite nor review here simply for the sake of brevity.

1 Basic Definitions, Assumptions and Randomised Experiments

1.1 Treatment Effects: Definitions, Assumptions and Problems

In econometrics, one often wants to learn the *causal effect* of a variable on some other variable, be it a policy question or some mere 'cause and effect' question. Although, at first glance, the problem might look trivial, it can become tricky to talk about causality when the real cause is masked by several other events. In this chapter we will present the basic definitions and assumptions about the casual models; in the coming chapters you will learn the different ways of answering questions about causality. So this chapter is intended to set up the framework for the content of the rest of the book.

We start by assuming we have a variable D which causes variable Y to change. Our principal aim here is not to find the best fitting model for predicting Y or to analyse the covariance of Y; we are interested in the impact of this *treatment* D on the outcome of interest (which is Y). You might be interested in the total effect of D on Y, or in the effect of D on Y in a particular environment where other variables are held fixed (the so-called *ceteris paribus* case). In the latter case, we again have to distinguish carefully between conditional and partial effects. Variable Y could indicate an outcome later in life, e.g. employment status, earnings or wealth, and D could be the amount of education an individual has received, measured as 'years of schooling'. This setup acknowledges the literature on treatment evaluation, where $D \in \{0, 1\}$ is usually binary and indicates whether or not an individual received a particular treatment. Individuals with $D = 1$ will often be called *participants* or *treated*, while individuals with $D = 0$ are referred to as *non-participants* or *controls*. A treatment $D = 1$ could represent, for example, receiving a vaccine or a medical treatment, participating in an adult literacy training programme, participating in a public works scheme, attending private versus public secondary school, attending vocational versus academic secondary schooling, attending a university, etc. A treatment could also be a voucher (or receiving the entitlement to a voucher or even a conditional cash transfer) to attend a private school. Examples of this are the large conditional cash transfer programmes in several countries in Latin America. Certainly, D could also be a non-binary variable, perhaps representing different subjects of university degrees, or even a continuous variable such as subsidy payments, fees or tax policies.

Example 1.1 The Mexican programme PROGRESA, which has been running under the name Oportunidades since 2002, is a government social assistance programme that started in 1997. It was designed to alleviate poverty through the rise of human capital. It has been providing cash payments to families in exchange for regular school attendance, health clinic visits and also nutritional support, to encourage co-responsibility. There was a rigorous (pre-)selection of recipients based on geographical and socioeconomic factors, but at the end of 2006 around one-quarter of Mexico's population had participated in it. One might be interested to know how these cash payments to members, families or households had helped them, or whether there has been any positive impact to change their living conditions. These are quite usual questions that policy makers need to answer on a regular basis. One key feature of PROGRESA is its system of evaluation and statistical controls to ensure its effectiveness. For this reason and given its success, Oportunidades has recently become a role model for programmes instituted in many other countries, especially in Latin America and Africa.

Let us set up the statistical setting that we will use in this book. All variables will be treated as random. This is a notational convenience, but it does not exclude deterministic variables. As measure-theory will not help you much in understanding the econometrics discussed here, we assume that all these random variables are defined in a common probability space. The population of this probability space will often be the set of individuals, firms, households, classrooms, etc. of a certain country, province, district, etc. We are thinking not only of the observed values but of all possible values that the considered variable can take. Similarly, we are not doing finite population theory but thinking rather of a hyper-population; so one may think of a population containing infinitely many individuals from which individuals are sampled randomly (maybe organised in strata or blocks). Furthermore, unlike the situation where we discuss estimation problems, for the purpose of identification one typically starts from the idea of having an infinitely large sample. From here, one can obtain estimators for the joint distribution of all (observed) variables. But as samples are finite in practice, it is important to understand that you can obtain good estimators and reasonable inference only when putting both together: that is, a good identification strategy and good estimation methods. Upper-case letters will represent random variables or random vectors, whereas lower-case letters will represent (realised) numbers or vectors, or simply an unspecified argument over which we integrate.

In most chapters the main interest is first to identify the impact of D on Y from an infinitely large sample of independently sampled observations, and afterwards to estimate it. We will see that, in many situations, once the identification problem is solved, a natural estimator is immediately available (efficiency and further inference issues aside). We will also examine what might be estimated under different identifying assumptions. The empirical researcher has then to decide which set of assumptions is most adequate for the situation. Before we do so, we have to introduce some notation and definitions. This is done in the (statistically probably) 'ideal' situation of having real experimental data, such as in a laboratory.

1.1.1 What Is a Treatment Effect?

There are many ways to introduce the notation of treatment effects. Perhaps the simplest is to imagine two parallel universes, say A and B, containing the same, identical individuals. When in universe A, individual i is now exposed to treatment $D_i = 1$, while in B it is not ($D_i = 0$); then all resulting differences for individual i, say $Y_i^A - Y_i^B$, can be considered as a treatment effect.

Let us formalise this by thinking of data-generating processes. For this, let D and Y be scalar random variables (an extension to vector-valued cases will be discussed later). Also assume in the following setup that Y is observed for every individual, whether it be employment status, earnings of wealth, etc.[1] We want to study the following relationship:

$$Y_i = \varphi(D_i, X_i, U_i) \,, \tag{1.1}$$

where φ is an unknown (measurable) function and (X_i, U_i) are vectors of observed and unobserved characteristics, respectively. The dimension of (X_i, U_i) is not yet restricted. Both might be scalars or of higher dimension; one even might want to drop X_i if only unobserved characteristics matter. If we consider abilities or skills as unobserved characteristics, then U_i can be multidimensional. Nonetheless, inside one equation, all unobserved parts are typically summarised in a one-dimensional variable. When we impose conditions on φ or the distributions of X_i and U_i, this can become relevant. In Equation 1.1 we assume that there exists a common φ for the whole population so that the right-hand variables comprise all heterogeneity when generating outcome Y_i. In this case, U_i plays the same role as the so-called residuals or error terms in regression analysis.

Thinking more generally, not just of one individual i, is important to emphasise that (1.1) does not imply homogeneous (i.e. the same for all individuals) returns to D or X even though we skipped index i from φ. Function φ just describes a structural relationship among the variables, and is assumed to be *not* under the control of the individuals; nor is it chosen or manipulated by them. As an example, it can be a production function. In particular, φ describes the relationship between D and Y not only for the actually observed values, but it should also capture the change in outcome Y if we had changed D externally to some other value.

In fact, our interest is to learn about this function φ or some features of it, so that we can predict what would happen if we changed D exogenously (i.e. without asking i). This idea becomes more clear when we define the concept of potential outcome. Notationally we express the potential outcome as

$$Y_i^d = \varphi(d, X_i, U_i) \,, \tag{1.2}$$

which is the outcome that individual i would experience if (X_i, U_i) were held fixed but D_i were set externally to the value d (the so-called *treatment*). The point here is not to enforce the treatment $D_i = d$, but rather to highlight that we are not interested in a

[1] In contrast, a variable like wages would only be observed for those who are actually working. The case is slightly different for those who are not working; clearly it's a latent variable then. This might introduce a (possibly additional) selection problem.

φ that varies with the individual's decision to get treated. The case where changing D_i also has an impact on (X_i, U_i) will be discussed later.

Example 1.2 Let Y_i denote a person's wealth at the age of 50, and let D_i be a dummy indicating whether or not he was randomly selected for a programme promoting his education. Further, let X_i be his observable external (starting) conditions, which were not affected by D_i, and U_i his (remaining) unobserved abilities and facilities. Here, D_i was externally set to d when deciding about the kind of treatment. If we think of a D_i that can only take 0 and 1, then for two values $d = 1$ (he gets the treatment) and $d = 0$ (he doesn't get the treatment), the same individual can have the two different potential outcomes Y_i^1 and Y_i^0 respectively. But of course in reality we observe only one. We denote the realised outcome as Y_i.

This brings us to the notion of a *counterfactual exercise*: this simply means that you observe $Y_i = Y_i^d$ for the realised $d = D_i$ but use your model $\varphi(\cdot)$ to predict $Y_i^{d'}$ for a d' of your choice.

Example 1.3 Let Y_i be as before and let D_i be the dummy indicating whether person i graduated from a university or not. Further, let X_i and U_i be the external conditions as in Example 1.2. In practice, X_i and U_i may impact on D_i for several individuals i such that those who graduate from a different subpopulation from those who do not might hardly be comparable. Note that setting externally D_i to d is a theoretical exercise, it does not necessarily mean that we can effectively enforce a 'treatment' on individual i; rather, it allows us to predict how the individual would perform under treatment (or non-treatment), generating the potential outcomes Y_i^1 and Y_i^0. In reality, we only observe either Y_i^1 or Y_i^0 for each individual, calling it Y_i.

Notice that the relationship (1.2) is assumed on the individual level to be given an unchanged environment: only variation in D for individual i is considered, but not variation in D for other individuals which may impact on Y_i or might generate feedback cycles. We will formalise this assumption in Section 1.1.3. In this sense, the approach is more focused on a microeconometric effect: a policy that changes D for every individual or for a large number of individuals (like a large campaign to increase education or computer literacy) might change the entire equilibrium, and therefore function φ might change then, too. Such kinds of macro effects, displacement effects or general equilibrium effects are not considered here, though they have been receiving more and more attention in the treatment evaluation literature. Certainly, i could be cities, regions, counties or even states.[2] In this sense, the methods introduced here also apply to problems in macroeconomics.

[2] Card and Krueger (1994), for example, studied the impact of the increase of the minimum wage in 1992 in New Jersey.

Example 1.4 An example of changing φ could be observed when a large policy is providing employment or wage subsidies for unemployed workers. This may lower the labour market chances for individuals not eligible to such subsidies. This is known as *substitution* or *displacement effects*, and they are expected to change the entire labour market: the cost of labour decreases for the firms, the disutility from unemployment decreases for the workers, which in turn impacts on efficiency wages, search behaviour and the bargaining power of trade unions. In total, we are changing our function φ.

Let us get back to the question of causality in the microcosm and look at the different outcomes for an exogenous change in treatment from d to d'. The difference

$$Y_i^{d'} - Y_i^d$$

is obviously the individual treatment effect. It tells us how the realised outcome for the ith individual would change if we changed the treatment status. This turns out to be almost impossible to estimate or predict. Fortunately, most of the time we are more interested in either the expected treatment effect or an aggregate of treatment effects for many individuals. This brings us to the average treatment effect (ATE).

Example 1.5 As in the last two examples, let $D_i \in \{0, 1\}$ indicate whether or not person i graduated from university, and let Y_i denote their wealth at the age of 50. Then, $Y_i^1 - Y_i^0$ is the effect of university graduation on wealth for person i. It is the wealth obtained if this same individual had attended university minus the wealth this individual would have obtained without attending university. Notice that the 'same individual' is not equivalent to the *ceteris paribus* assumption in regression analysis. We explicitly want to allow for changes in other variables if they were caused by the university graduation. While this is doubtless of intimate interest for this particular person, politicians might be more interested in the gain in wealth on average or for some parts of the population.

Sometimes we want to consider situations explicitly where we are interested in the effects of two (or more) treatment variables. We could then consider D to be vector-valued. Yet, it is useful to use two different symbols for the two treatment variables, say D and X (subsumed in the vector of observable characteristics), for two reasons. The first reason is that we may sometimes have some treatment variables D that are endogenous, i.e. caused by U, whereas the other treatment variables are considered exogenous. Therefore, we distinguish D from the other variables since dealing with the endogeneity of D will require more attention. A second, unrelated reason is that we sometimes like to make it explicit that we are mainly interested in the impacts of *changes* in D by external intervention while keeping X fixed. This is the known *ceteris paribus* analogue, and in the treatment effect literature is typically referred to as a *partial* or *direct effect* of D on Y.

Example 1.6 Let D_i indicate whether individual i attended private or public secondary school, whereas X_i indicates whether the individual afterwards went to university or not. Here, we might be interested in that part of the effect of private versus public school on wealth that is not channelled via university attendance. Clearly, attending private or public school (D) is likely to have an effect on the likelihood to visit a university (X), which in turn is going to affect wealth. But one might instead be interested in a potential direct effect of D on wealth, even if university attendance is externally fixed. From this example we can easily see that it depends heavily on the question of interest, i.e. how the treatment parameter is defined.

To notationally clarify the difference to the above situation, let us define

$$Y_{i,x}^{d} = \varphi(d, x, U_i) \; .$$

The function φ is still the same as before, so the only difference is that D and X are both thought of as (separate) arguments that one might want to set externally. Then the *partial* or *direct effect* (i.e. the effect not channelled via university attendance in Example 1.6) of D (public versus private school) is

$$Y_{i,x}^{d''} - Y_{i,x}^{d'} \; .$$

That is, $Y_{i,0}^{1} - Y_{i,0}^{0}$ in Example 1.6 is the partial effect of private/public when university attendance is set to zero, whereas $Y_{i,1}^{1} - Y_{i,1}^{0}$ is the effect when university attendance is fixed at one (by external intervention). In contrast, the total effect of private versus public secondary school is

$$Y_{i}^{d''} - Y_{i}^{d'} \; .$$

Hence, the reason for using two different symbols for D and X is to emphasise that one is interested in the effects of changes in D while keeping X fixed or not. Sometimes such partial effects can be obtained simply by conditioning X, and sometimes more sophisticated approaches are necessary, as will be seen later.

Example 1.7 Consider the Mincer earnings functions in labour economics, which are often used to estimate the returns to education. To determine them, in many empirical studies log wages are regressed on the job experience, years of schooling and a measure of ability (measured in early childhood, if available). The reasoning is that all these are important determinants of wages. We are not so interested in the effects of ability on wages, and merely include ability in the regression to deal with the selection problem discussed later on. The *ceteris paribus* analysis examines, hypothetically, how wages would change if years of schooling (D) were changed while experience (X) remained fixed. Since on-the-job experience usually accumulates after the completion of education, schooling (D) may have different effects: one plausible possibility is that schooling affects the probability and duration of unemployment or repeated unemployment, which reduces the accumulation of job experience. Schooling outcomes

may also affect the time out of the labour force, which also reduces job experience. In some countries it may decrease the time spent in prison. Hence, D affects Y indirectly via X. Another possibility is that years of schooling are likely to have a direct positive effect on wages. Thus, by including X in the regression, we control for the indirect effect and measure only the direct effect of schooling. So, including X in the regression may or may not be a good strategy, depending on what we are trying to identify. Sometimes we want to identify only the total effect, but not the direct effect, and sometimes vice versa.

We are interested in *non-parametric identification* of φ or some features of it. Non-parametric identification basically means that there is no further model than Equation 1.1 without further specification of φ. Thus, the identification will be mainly based on assumptions relating to the causality structure, which in practice have to be based on economic theory. In contrast, most econometric textbooks start by assuming a linear model of the type (keeping our notation of variables)

$$Y_i = \alpha + D_i\beta + X_i\gamma + U_i \tag{1.3}$$

to discuss identification and estimation of β under certain restrictions like

$$E[U_i|D_i, X_i] = 0\,. \tag{1.4}$$

In other words, they identify the parameters of (1.3), which coincide with the question of interest only if their model is correctly specified. The statistics literature typically discusses the correct interpretation of the parameter whatever the true underlying data-generating process may be, to relate it to the question of interest afterwards. Doubtless, the latter approach is safer, but it might answer the question of interest only unsatisfactorily. However, since the assumption of linearity is almost always an assumption made for convenience but not based on sound economic theory, it is more insightful to discuss what can be identified under which restrictions without imposing a functional form on φ; it might be linear, quadratic or any other form; it need not even be continuous, differentiable or monotonic.

For identification, we will nonetheless have to impose certain restrictions, which are usually much weaker than (1.3), (1.4). Such restrictions often come in the form of differentiability and continuity restrictions on φ (also called smoothness restrictions). There is a large and still growing literature which attempts to find the weakest assumptions under which certain objects can be identified. The function φ is *non-parametrically identified* if we can determine it exactly from an infinitely large sample. Suppose that we have infinitely many observations, so that we effectively know the joint distribution of Y, D and X. The function φ, or some feature of it, is non-parametrically identified if no other function could have generated the same distribution. Or, putting it the other way around, it is not identified if two different functions, say φ and $\tilde{\varphi}$, could generate the same joint distribution of the observed variables. A consequence of the lack of an explicit (parametric) model or function φ is that it is now identified only in some regions but e.g. not outside the support of the observations.

1.1.2 Formal Definitions: ATE, ATET, ATEN

In this section we will formalise different treatment effects 'on average'. We focus on a binary treatment $D \in \{0, 1\}$ case, because it helps a lot to understand the main issues of identification without further complexities; later on we will also discuss some extensions. Recall the university graduation of Example 1.5. There we wanted to estimate the expected wealth effect of attending university for a person randomly drawn from the population, namely

$$E[Y^1 - Y^0].$$

Notice that the expectation operator has the same meaning as averaging over all individuals i of the population of interest. To put things in perspective, let's consider a city in which we send everyone to universities (of course, assuming we have the authority to do so). Somewhat later, we observe their income and take the average for calculating $E[Y^1]$. Now let's imagine we travel back to the past and keep everybody away from university. Again we observe their income and calculate the average to get $E[Y^0]$. Since expectation is linear, this difference is exactly the same as we mentioned before. This is known as an *average treatment effect*, or ATE for short. In reality we cannot send an individual to both states. Some of them will go to university and some won't. So, for the same individual, we cannot observe his outcomes in both states, we observe only one. We can interpret the average causal effect of attending university in two ways: it is the expected effect on an individual randomly drawn from the population and at the same time, and it is the change in the average outcome if D were changed from 0 to 1 for every individual, provided that no general equilibrium effect occurs.

In reality, some of the individuals will go to university and some won't. What we can do is then to take the average of those who attended the university and those who didn't, respectively. To compare the average wealth of those who attended ($D_i = 1$) with those who didn't ($D_i = 0$), we could look at

$$E[Y^1|D = 1] - E[Y^0|D = 0].$$

It is important to see that conceptually this is totally different from an average treatment effect. First of all, we didn't do anything; it was they who decided to go to university or not. This creates the two groups different in many ways. Particularly they might differ in observed and unobserved characteristics. This difference is most apparent when examining the effect only for those who actually did attend universities. This is the so-called *average treatment effect on the treated* (ATET) and it is defined as

$$E[Y^1 - Y^0|D = 1].$$

Again, you do the similar thought experiment as above but not for the whole city; rather, just for those who actually would have attended universities anyhow. This is often of particular interest in a policy evaluation context, where it may be more informative to know how the programme affected those who actually participated in it than how it might have affected others. Here it is even more obvious that simply comparing the observed outcomes of those who did attend university with those who did not will usually not provide a consistent estimate of the ATET. Already intuition tells us that these