Causal estimation and causal inference.

Eduardo Fé[1]

April 30, 2019

[1]Department of Social Statistics, University of Manchester. email: eduardo.fe@manchester.ac.uk.

Chapter 1

Causal estimation and causal inference.

1.1 Introduction.

The origins of statistics as a formal discipline are tightly intertwined with modern states' efforts to objectively quantify their wealth and health as an attempt to build a solid foundation for policy making. Statistics has, at least implicitly, sought to answer questions about the causes, drivers and mechanisms of the world: from whether 'miasmas' were ultimately responsible for cholera outbreaks in Victorian Soho, to whether an aggressive interest rate policy might have shortened the life-span (or moderated the political consequences) of the 2008 financial crisis.

It is thus paradoxical that, on its own, the traditional language of statistics is not designed to provide answers to the causal questions that plague the agendas of governments, business, academics and the general public. The traditional *focus* of statistics has been the development of techniques for data 'reduction' or estimation and finding parsimonious equations to model the joint behaviour of sets of variables. Good estimation techniques are critical for any statistical analysis but, on their own, they lack of any causal interpretation, even when their mathematical and probabilistic foundations are well established and understood. Statistical models, on the other hand, might reveal, at best, no more than a researchers' own views on the underlying causal mechanisms, despite their mathematical foundations, their sophistication or the amount of data available.

Researchers have in recent times expanded the statistical tool kit with a new methodology specifically designed to uncover the working causal mechanisms of the world. Foundational contributions during the twentieth century, and particularly since the 1970s, have introduced and standardised the formal statistical language of causality. These methods have, in particular, led to a silent 'causal revolution' over the last four decades. During this revolution, researchers' capabilities to undertake data-based analysis of causal ques-

tions and, critically, understand the premises and limitations of these studies, has been substantially magnified. Some of the foundations and methods of this silent revolution are the object of the present chapter.

The need for a new causal lexicon is well illustrated by Simpson's paradox, which refers to inconsistencies in the sign of statistical associations when the focus of analyses shifts from the population to its constituent subpopulations. Far from being a theoretical curiosum, this paradox is arises frequently in practice, as illustrated by the seminal study of Bickel, Hammel, and O'Connell (1975). These authors observe statistically significant differences by gender in admission rates for graduate study at UC Berkeley for the autumn 1973 quarter. Out of 44% of 8442 male applicants were admitted, in comparison to 35% of 4321 female applicants. At face value, this finding is morally and legally troubling. The paradox arises when Bickel et al. (1975) examine their data at departmental level. That analysis suggests that, instead, there is a small but statistically significant bias in favour of women, which contradicts the results of the aggregated analysis. How can this result emerge? Although the pool data suggests the existence of significant discrimination, this conclusion is warranted only under the assumption that the distribution of applications across genders are qualitatively the same and, in particular, that the proportion of women applying to any one specific department equals the proportion of men applying to that same department. This, however, is not the case. Bickel et al. (1975) find that women tend to apply to graduate departments that are more crowed and were, therefore, it is difficult for applicants of either sex to enter. Within these departments data showed small but statistically significant bias in favour of women. Bickel et al. (1975) go on to argue that:

'*The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.* '.

Simpson's paradox does not imply that the estimates derived from the pooled data were wrong (they were statistically unbiased and consistent for the *estimable parameters*). Rather, it emphasises the critical features of any causal analysis. The first feature is the common need to draw information external to the primary analysis in order to endow any conclusions with a causal interpretation (in the case of Bickel et al. (1975), this information came from the intuition that led those authors to investigate if the distribution of applications by department might matter for admission). Though important, however, this feature is case specific, and thus this chapter will have little to say about

it (the interested reader can see excellent monographs by Dunning (2012) and Diamond and Robinson (2011)). The second, and most fundamental aspect, however, is that the assumptions underpinning the analysis drive the conclusions obtained by the researcher. This latter aspect will be the focus of much of the discussion that follows.

When it comes to articulating a general framework within which to develop statistical analysis of causes and effects, we find two closely related schools of thought. On the one hand, we find the potential outcomes framework, developed from the seminal papers by Rubin (1974) and Holland (1986). This framework enables researchers to clearly define the parameter being estimated, as well as being very explicit about the various assumption underpinning any causal analysis. Researchers can then try to understand the suitability of each specific assumption to the analysis under consideration. The discussion in this chapter is framed within the potential outcomes framework, which is introduced in the next section. The second school of thought falls within the framework of the graphical theory of causality by Judea Pearl and his collaborators (Balke and Pearl (1997); Pearl (2009b); Pearl, 2009a). The main characteristic of this approach is the use of directed acyclic graphs to build up a model describing the main casual relationships underlying the problem at hand. This inverts the flow of work with respect to the potential outcomes framework. Pearl's framework first establishes the many potential causal relationships underlying the data and, subsequently, estimation is tackled within a classical or Bayesian framework. In the potential outcomes framework, the parameter and assumptions enabling estimation are defined a priori, with researchers having to subsequently explore the suitability of each assumption for the specific problem at hand. Pearl's framework is thus parametric in nature, whereas the potential outcomes framework is naturally non-parametric. Both approaches are, however, complementary, and there is a number of subtle commonalities to these approaches. For details, the interested reader can refer to Morgan and Winship (2014) and Pearl, 2009b for an excellent introduction.

1.2 The potential outcomes framework.

The discussion in this chapter relies on the potential outcomes framework, developed in Rubin (1974) and Holland (1986) (see also Rubin, 1990a), which provides a unifying context to study both randomized experiments and observational studies. Throughout the chapter we will focus on those studies where there are two 'experimental' conditions, which we will call 'treatment' and 'control'.

There are three constituent ingredients in the potential outcomes framework. First, we need to define the 'unit' of analysis, which will be an entity (a person, firm, classroom, cell, etcetera) with measurable traits *at a given point in time* (and so, the same entity at different moments constitutes different units). Second, we need to define an active treatment, which is an action or manipulation that might affect or not a particular unit.

If a unit does not receive the active treatment, then the unit must have received the 'control' treatment. We will often also say that units are allocated to the treatment or control groups to mean whether or not a unit received the active treatment. Third, associated with each unit, there are two potential outcomes at a *future* point in time, which specify the level of some outcome if the active treatment is received by a given unit and the level of the outcome if the control treatment is received by that unit.

More formally, let T indicate the treatment status of a unit, so that T equals 1 if the unit has received the active treatment and T equals 0 if the unit has received the control treatment. Following RUBIN74, $Y_i(t)$ denotes the potential outcome for unit $i = 1, \ldots, N$ under treatment $t \in \{0, 1\}$. Each unit reveals only one of these two potential outcomes. In particular,

$$Y_i = Y_i(1) \cdot T_i + Y_i(0) \cdot (1 - T_i) \tag{2.1}$$

This potential outcomes framework can be traced back to the works of Jerzy Neyman's masters dissertation '*Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principestranslated*' (1923; partly translated in Neyman, Dabrowska, and Speed, 1990; see also Neyman, 1934 and the commentary in Rubin, 1990b), who used it in the context of randomized experiments. However it was Rubin (1974) who exploited this notation to provide a unifying framework for the study of both randomized experiments and observational studies.

The causal effect of the treatment T results from comparing $Y(1)$ and $Y(0)$. For example, inferences could be based on the difference in levels $Y(1) - Y(0)$; it could be based on a ratio of levels $Y(1)/Y(0)$; it could be based on a meaningful transformation such as $\log Y(1) - \log Y(0)$. The important aspect is that the definition of the causal effect will depend on the potential outcomes only, and not on which potential outcome is observed. Furthermore, the causal effect involves the comparison of potential outcomes of the same unit at the same time after receiving the treatment. In particular, the treatment effect is not defined by comparing potential outcomes at different points in time (e.g. before and after treatment).

The potential outcomes framework is readily suited to formally describe well formulated causal questions. Perhaps more critically, the potential outcomes framework helps researchers to detect poorly formulated causal statements and think about better or more precise ways formulate these statement. For instance, consider the statement 'She had low levels of cholesterol in blood because she consumed statins every day'. The treatment in this example is clear, the consumption of statins, a medicine which is specifically prescribed by doctors at specific dosages to lower cholesterol in blood. The outcome variable is well defined and there is a temporal sequence of events, with treatments following outcomes. The alternative to the treatment not to consume statins every day; this is

somehow imprecise and more valuable conclusions could be drawn by perhaps specifying if the alternative would be an irregular intake of statins or no intake at all. Consider, in contrast, this other assertion '*Austerity following the 2008 financial crises caused social unrest in the United Kingdom in 2016*'. The statement describes a treatment (austerity), a logic sequence of events and an outcome. However, austerity is an elusive concept. It is typically associated with funding cuts, but the dosage of this treatment is likely to vary across units (a point to which we return when discussing the Stable Unit Treatment Value Assumption -SUTVA- below). The alternative to austerity is not clear either. It could range from a funding cuts of different magnitude, to no cuts or even to raises in public funding. Similarly, social unrest is not a well defined outcome, being an multifaceted construct which could include concepts as diverse as crime, riots, unemployment, demonstrations, protest voting, and so on. Thus, the second is an ill defined causal statement.

Besides helping researchers to clearly *define* valid causal questions, the potential outcomes framework emphasises the pervasive problem of selection into treatment. Given any one unit, both potential outcomes cannot be observed simultaneously. This implies that, although the potential outcomes framework helps to clearly defined the unit level causal effect, the latter is not identified or estimable from the data -there is a problem of *missing data*. Since a unit can only provide one potential outcome, we will need to collect data from many units in order to have sufficient information about what happens to the outcome under analysis following treatment and in the absence of treatment. The use of information from different units, however, creates additional problems for both the definition and estimation of causal effects. Consider the problem of estimating the effect of statins on cholesterol using data from two individuals. Then there are four treatment *pairs*, (t_1, t_2) (either both individuals have statins, or nobody has statins, or only one individual has statins) and thus eight potential outcomes $Y_i(t_1, t_2)$ (four per individual). In this example we can define causal effects based on six meaningful comparisons such as, for instance, a comparison of $Y_1(1, 0)$ with $Y_1(0, 0)$ or $Y_2(0, 1)$.

To define a causal parameter when multiple units are considered we need to first specify if treatments are comparable across units and, second, whether or not the treatment status of a unit might affect the outcomes of other units. In this regard, most published literature in causal inference rests on the Stable Unit Treatment Value Assumption (SUTVA), (Rubin, 1980) defined as,

Assumption 1 (SUTVA) There is no interference between units and there are no versions of treatments leading to "technical errors"

SUTVA is already implicit in the notation in equation (2.1). SUTVA has two components. The first and most widely acknowledge is 'no interference' which implies that outcomes do not depend on the treatments that other units received. No interference was a central concern for the agricultural experiments which were the focus of research by

early pioneers of causal inference (e.g. Neyman et al., 1935; Fisher (1935); Rubin, 1990b; Cox, 2009). In those settings, guard rows in experimental plots and other physical devices could be deployed to mitigate cross-contamination of plots. In the social sciences, however, no interference can often be a troublesome assumption because 'guards' are generally unavailable. A classic example of this problem is immunisation, where inoculation of vaccines to substantial sets of the population can generate positive externalities for the non-vaccinated. As another example, consider the evaluation of an education policy or intervention focusing on students as a unit of analysis might be problematic because of potential spillovers between students within the same school or academic year. Often, however, changing the definition of the unit of analysis may render SUTVA less controversial. In the context of schools, this can be done by defining classrooms or schools themselves as the unit of analysis.

The second component in SUTVA is the absence of 'technical errors' by which Rubin means no hidden variation in treatments (Rubin attributes the term 'technical error' to Cox, 1958). Consider an experimental study of the effectiveness of statins on blood cholesterol. The idea here is that all participants taking the same dosage of the medicine are receiving identical amount of the active ingredient; we would thus rule out systematic latent under or over doses due to, for example, a failure in production at the factory where the statins are made (note, however, that Cox's original definition of 'technical errors' allows for random, non-systematic variation in dosage). This second dimension of SUTVA is very problematic in observational studies. Consider, for example, research on the effects of retirement on consumption (Battistin et al. (2009), Banks et al. (1998)) or health (Charles (2004); Neuman (2008); Coe and Zamarro (2008)). Most of these studies implicitly assume that the treatment (retirement from the workforce) is not subject to hidden variations. It is, however, difficult to objectively argue that this is the case. Variation in treatment in these studies can be due to differences in pre-retirement wages, post retirement pensions, pre-retirement working conditions, among otheres. This further suggests that the credibility of SUTVA might hinge on our target population and the definition of the unit of analysis.

We will maintain SUTVA throughout the discussion. However, it is important to bear in mind that this is a substantive assumption and, as such, it must be properly justified in research studies. From this perspective, the availability of pre-treatment information can be critical in order to justify the no-interference and no-hidden variation premises underlying SUTVA. Formally, the implications of SUTVA from a notational perspective consider a number of units $i = 1, \ldots, N$. Then no hidden variation implies that we can confidently define a single treatment indicator T which takes value 1 if the treatment is assigned to a unit (0 otherwise). In general, each unit's potential outcomes is given by $Y_i(t_1, t_2, \ldots, t_N)$ where t_i is the treatment status of unit $i = 1, \ldots, N$. Under no-interference, however, we can define each unit's potential outcome as a function of its own

Unit	Potential outcomes		Causal effect	Assignment A		Assignment B	
	Y(0)	Y(1)	Y(1)-Y(0)	T	Observed	T	Observed
Patient 1	135	136	1	0	135	1	136
Patient 2	220	156	-64	1	156	1	156
Patient 3	247	206	-41	1	206	0	247
Patient 4	156	149	-7	0	156	1	149
Patient 5	226	198	-28	1	198	0	226
Patient 6	127	127	0	0	127	0	127
Average	185.17	162					
True average effect			-23.17				
Estimated average effect					47.34		-39.75

Table 1.1: The importance of the assignment mechanism. The outcome in the table is Low Density Lipoprotein in blood, in milligrams per decilitre. Assignment A allocates statins to patients who would benefit the most. Assignment B allocates the treatment at random.

treatment status, $Y_i(t_i)$.

1.3 The assignment mechanism.

Despite its strength, SUTVA does nothing to ameliorate the fundamental problem of causal inference, which is missing data on one potential outcome per unit. Indeed, this problem is insurmountable without the incorporation of carefully chosen assumptions by the investigator. In this respect, the most critical assumptions will be those contributing to (or shaping our) understanding of the process that determined which units were allocated to the active treatment or control treatment. Within the potential outcomes framework, that process is known as the *assignment mechanism*, which is a characterisation if the likelihood of each feasible distributions of units into treatment and control groups.

To see how critical our understanding of the assignment mechanism is for causal inference consider two study protocols of the effects of statins on low-density lipoprotein (LDL) cholesterol in blood on six patients. Doctor B decides to allocate statins to patients at random, by flipping a coin. Doctor A decides that B's protocol is unethical and, instead, he screens the participants and relies on his expertise to detect those individuals who are most likely to benefit from the statins. A prescribes statins only to this group. Both doctors will evaluate the success of statins by comparing the average outcomes between treated and non-treated individuals. The situation is described in Table 1.1. In this example, statins are beneficial in general, particularly for those individuals with very high LDL levels (above 200 milligrams per decilitre). In this sample of six patients, statins lead to a reduction of 23 milligrams per deciliter in LDL levels. Under protocol B, the estimated average effect of statins is also negative, but at -39.75 it overestimates the true effect. Under the more ethical protocol A, however, the estimated effect is positive. Taken

at face value, one would thus conclude that statins are detrimental for health. However, this conclusion would neglect the fact that only those individuals who benefited the most from statins actually received the medicine. In fact, by allocating the medicine to these individuals, we have achieved an average drop of LDL of 44 milligrams per decilitre.

The standard paradigm of causal inference starts by introducing two structural assumptions on the assignment mechanism. First, it is standard practice to assume that treatment assignment is *individualistic*, that is a unit's treatment status does not depend on the outcomes and assignments of other units. This assumption would be violated when considering the effect of retirement decisions on, for example, household expenditure. In that case the retirement of one unit in a household is likely to take into consideration the distribution of expenditure in the full household as well as the retirement status of any partner (which will ultimately determine household income). A second structural assumption is that the assignment mechanisms are '*probabilistic*'. This means that any unit has a non-zero probability of being allocated to the treatment or control group. This is a technical assumption which can be seen as ruling out uninformative units in any sample.

The defining assumption in causal inference, however, will refer to whether the assignment is confounded or unconfounded. The assignment mechanism is unconfounded if it does not depend on units' potential outcomes, so that for any unit $T \perp Y(0), Y(1)|\mathbf{X}$ (for some k-vector of pre-treatment variables $\mathbf{X}$, which might be empty). This would eliminate situations where units self-select into treatment on the basis of perceived advantages or disadvantages derived from the treatment. For example, when studying the effect of retirement on health, unconfoundedness would imply that individuals facing retirement do not take into consideration the potential advantages of leaving the workforce for their own health. When an assignment mechanism is unconfounded, individualistic and probabilistic, it is referred to as an strongly ignorable treatment assignment.

Having introduced the idea of unconfoundedness, we can finally offer a taxonomy of assignment mechanisms depending on whether or not the researcher has full knowledge *and control* of the assignment mechanism. For practical purposes, we can distinguish between:

- A *classical randomized experiment* is an strongly ignorable treatment assignment with a known form which is controlled by the experimenter.
- Observational studies where the assignment mechanism is not controlled or known by the experimenter.
  - A *regular assignment mechanism* is an observational study with a strongly ignorable treatment assignment
  - *Irregular assignment mechanism.* For our purposes, this will be observational studies where *assignment* to treatment is *unconfounded* but the actual reception of treatment is confounded.

The focus of this chapter will be on the classical randomised experiment and observational studies with irregular assignment mechanisms. Regular assignment mechanisms are common in practice, but they have the peculiarity that, at an elementary level, most of the techniques applicable to randomised experiments can be extrapolated to that context, with a number of variations to ensure that unconfoundedness is satisfied in practice. The interested reader will find excellent discussions of regular assignment mechanisms in Imbens and Rubin (2015), Angrist and Pischke (2008), Morgan and Winship (2014) among others.

1.4 Neyman, Fisher, Finite samples and Populations.

The discussion in this chapter will take a (super)population perspective by which data in a sample is the result of a random draw from a population whose size tends towards infinity. Within this perspective, Neyman's framework for causal inference will provide the unifying framework for the discussion. These two characteristics, however, provide just one among several competing perspectives on the problem of causal inference.

In Neyman's causal framework, the focus is on the average effect of the treatment on the population, understood as the difference in the average level of outcome when everybody receives the active treatment, $\bar{Y}(1)$ and the average level of outcome when everybody receives the control treatment, $\bar{Y}(0)$. More precisely, the parameter of interest is

$$\tau = E\Big[\bar{Y}(1) - \bar{Y}(0)\Big] \tag{4.1}$$

From this perspective, an ineffective active treatment might still have positive or negative effects for the levels of Y in the population but, on average, the positive and negative effects cancel each other out, resulting in an overall null average effect. The superpopulation setting corresponds to the traditional framework in the social sciences, where researchers are typically interested in inferring the effect of a policy for a whole a region or country when only a small sample might be available for the analysis. In this superpopulation setting, unit level potential outcomes are fixed. However, random sampling induces a distribution on the unit level potential outcomes (which now become random variables), the unit level treatment effect and the average unit level treatment effect. Within this superpopulation setting, most statistical result relating the the interpretation and characteristic of estimators will rely on asymptotic arguments based on the delta-method, laws of large numbers and the central limit theorem.

An alternative to the superpopulation setting is a finite population framework with a fixed number of units, N. In this setting potential outcomes are fixed and variation in outcomes is driven by variation in treatment assignment. In this setting, the parameter

of interest is, therefore

$$\tau_{fp} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0) = \bar{Y}(1) - \bar{Y}(0) \tag{4.2}$$

In this setting, it is often possible to characterise or infer the details of the assignment mechanism. As a result, one can often stablish small sample properties of estimators (such as unbiasedness in finite samples) and the distribution of tests of hypothesis. In the finite population perspective the population and the sample are equivalent concepts. Thus, this perspective emphasises that results of any statistical analysis are circumscribed to the realm of the sample. This might seem restrictive, but it endows analyses with greater credibility.

The main limitation in our discussion comes, however, from obviating Fisher's randomization inference approach to causality. As in the finite population approach, the Fisherian (Fisher, 1935) framework assumes that a unit's potential outcomes are fixed. However, in the latter setting an ineffective active treatment means that no unit sees its level of outcome altered upon receiving the treatment. That is, Fisher defines a 'sharp' null hypothesis by which $Y_i(1) = Y_i(0)$ for all i. The implication of this is that, under the sharp null hypothesis, we can automatically infer the full distribution of potential outcomes for the observed data. Furthermore, given any test of the 'sharp' null hypothesis, we can obtain its exact or approximately exact distribution from the randomization of units across treatment (the randomization distribution). Fisher's approach enjoys the characteristic of being fully non-parametric because it does not require any model for the distribution of outcomes, whereas the assumptions required to characterise the randomization distribution of any test tend to be minimal. From the perspective of clinical trials and lab experiments, where researchers have full control of the assignment mechanism, Fisher's randomization inference methods provides a natural framework for analysis. Furthermore, this framework is readily suited to test hypothesis about non-standard parameters, such as quantiles of distributions of outcomes. A potential limitation of Fisher's randomization inference approach is that point estimation comes as a by-product of hypothesis testing. Therefore researchers need to invert tests in order to obtain point estimates. This type of Lehmann-Hodges estimator (Hodges and Lehmann, 1963) tends to rely on somehow strong assumptions (such as linearity and homogeneity of treatment effects) and their calculation requires numerical and simulation methods, which can be computer intensive (a relative cost, given the power of modern computers). On this points, see Ho and Imai (2006); Rosenbaum (1996); Rosenbaum and Imbens (2005), Rosenbaum (2010) and references therein.

1.5 Unconfounding: Classical Randomised Experiments.

Most researchers agree that randomized experiments, when feasible, are very strongly suited to draw causal inferences, although, as noted by Senn (2013), they are not sufficient for answering all 'our evidential needs'. Although traditionally used in biological, agricultural and medical settings, experiments have now become popular in many other fields of inquiry, including psychology, education, economics and political sciences. Their strength arises from their design. In these experiments the active and control treatments are allocated at random. When properly designed, the observable and unobservable characteristics of the units do not influence the assignment to treatments and, as a result, any variation in observable outcomes can be attributed to the treatment. Randomized experiments also provide a useful benchmark against which observational studies can be compared in order to gain a better and clearer understanding of the causal information that observational studies can provide. Furthermore, as we will see in a later section, many of the estimation techniques in these settings can be used when estimating causal effects in observational studies. Despite of their advantages, consensus about the benefits of randomized experiments is not unanimous. Indeed, many influential researchers have vehemently argued against randomized experiments. The range of criticisms are various, but are well summarised in the 2018 symposium 'Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue' in the journal Social Science and Medicine. For criticisms against randomized experiments, see, in particular, Deaton and Cartwright (2018), Cook (2018), Pearl (2018); arguments in favour of randomized experiments are provided by Imbens (2018) and Raudenbush (2018) among others.

Formally, classical randomized experiments are defined by an assignment mechanism which is individualistic, probabilistic, unconfounded and has a known functional form that is controlled by the researcher. In this chapter, we focus on the case of *Completely Randomized Experiments* where N_t out of N units are randomly assigned to receive the active treatment, with the remaining $N_c = N - N_t$ units receiving the control treatment. Note, however, that there is a non-zero probability that this type of assignment might result in an uninformative distribution of units across experimental groups. For example, in a completely randomized experiment one could end up allocating all units of a specific subpopulation (e.g. men) to one of the treatment groups. Alternative assignment mechanisms can be considered that limit or rule out the likelihood of such uninformative assignments. In a stratified experiment, for example, random allocation to treatments occurs after units are separated in strata in accordance to the value of a *relevant* variable that might have bearing on the levels of the outcome under exploration. Alternatively, one can pair units which are close to each other in regards to the values of one or several characteristics and then randomly allocate one unit in each pair to the active treatment

-with the remaining unit being allocated to the control treatment. Paired and stratified experiments are, however, beyond the scope of this survey. The interested reader can find a detailed discussion in Imbens and Rubin, 2015.

As mentioned above the focus is on Neyman's approach to classical randomized experiments, and so we are interested in understanding the average differences that would be observable between two counterfactual scenarios: one where all the population receives the active treatment and one where all the population receives the control treatment. This perspective is, at least conceptually, less restrictive than the Fisherian sharp null hypothesis approach. In the latter, an ineffective active treatment implies a null effect for all units. In Neyman's approach, an ineffective treatment implies an average 0 effect, even though the unit level treatment effect might be non-zero across units.

Working within the population perspective, our sample is understood as the result of a random draw from an population of infinite size. This implies that, although the potential outcomes are considered to be fixed at unit level, random sampling induces distributions for the potential outcomes, the unit level treatment effect and the average of the unit level treatment effects. Specifically, then, the parameter of interest in our Neyman approach to classical randomised experiments is,

$$\tau = E(Y_i(1)) - E(Y_i(0)) = E\big[Y_i(1) - Y_i(0)\big] \tag{5.1}$$

where expectations are taken with respect to the distribution of potential outcomes in the super-population. Note that the above notation already incorporates SUTVA. Also, the above pair of equalities specifies the three different parameters that were of original concern for Neyman, namely the population average outcome under the active and control treatments, $E(Y(1))$ and $E(Y(0))$ respectively, and the population average unit treatment effect, $E\big[Y_i(1) - Y_i(0)\big]$. The equality of these parameters is the result of the assumptions underlying Classical Randomized Experiments and SUTVA.

Given a finite sample of N observations from the superpopulation, we can estimate τ by means of the finite sample average treatment effect,

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \cdot T_i - \frac{1}{N_0}\sum_{i=1}^{N} Y_i \cdot (1 - T_i). \tag{5.2}$$

This estimator is unbiased for τ conditional on the distribution of the potential outcomes in the superpopulation. Under random sampling, the variance of the estimator conditional on the distribution of potential outcomes equals the sum of the variances of the treatment and control groups in the population,

$$V(\hat{\tau}) = \frac{\sigma_1}{N_1} + \frac{\sigma_0}{N_0} \tag{5.3}$$

where $\sigma_s = V(Y_i(t)) = E\big[\big(Y_i(t) - E(Y_i(t))\big)^2\big]$ for $t = 0, 1$. Let $R_{1,i} = Y_i \cdot T_i$ and $R_{0,i} = Y_i \cdot (1 - T_i)$. Then a conditionally unbiased estimator of the variance is given by

$$\hat{V}(\hat{\tau}) = \frac{s_1}{N_1} + \frac{s_0}{N_0} \tag{5.4}$$

$$s_1 = \frac{1}{N_1 - 1}\sum_{i=1}^{N}(R_{1,i} - \bar{R}_{1,i})^2 \text{ and } s_0 = \frac{1}{N_0 - 1}\sum_{i=1}^{N}(R_{0,i} - \bar{R}_{0,i})^2 \tag{5.5}$$

$$\bar{R}_{t,i} = \frac{1}{N_t}\sum_{i=1}^{N} R_{t,i} \tag{5.6}$$

Interestingly, under random sampling, the expected value of this estimator equals the the variance of the unit level treatment effect in the superpopulation. For large samples, then we can construct confidence intervals with nominal coverage $1 - \alpha$ based on the normal approximation,

$$CI_\alpha = \left(\hat{\tau} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}\ ,\ \hat{\tau} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}\right) \tag{5.7}$$

where $z_{\alpha/2}$ is a quantile from the normal distribution. Naturally, we can use the estimator of the variance to construct tests of hypotheses about the average treatment effect and, in particular, its statistical significance. Specifically, we can test the null hypothesis of no significant effect of the active treatment on the basis of the ratio,

$$\nu = \frac{\hat{\tau}}{\sqrt{\hat{V}(\hat{\tau})}} \tag{5.8}$$

which has an asymptotic normal distribution. Therefore, give a value of the above ratio, ν^*, we would reject the null hypothesis of no statistical effect of the active treatment at 5% significance level if $|\nu^*| > 1.96$.

Regression perspective.

It is straightforward to extend the preceding framework to situation when researchers believe that the active treatment might have a differential effect across subsets of the population defined by specific values of a relatively small set of discrete pre-treatment variables. Neyman's estimator can be calculated across the different subpopulations. Subsequently, a weighted average of these estimates provides an unbiased estimator of the overall treatment effect. When one suspects that the outcome under consideration might be driven by a large set of continuous and discrete variables, however, one might find a dearth of data within certain sub-categories or that only treated or untreated units have been observed for these sub-categories. In these settings one might want to consider a regression analysis of the treatment effect. If the pre-treatment information is hihgly

predictive of the levels of the outcome, then the regression approach might result in inferences that are more precise than those obtained with Neyman's estimator. Interestingly, however, in the context of randomized experiments, treatment is independent of other covariates. As a result, the properties of the estimator of the treatment effect do not hinge on how well the overall regression model captures the true conditional mean in the population.

The traditional approach to regression (e.g. Wooldridge, 2010; Angrist and Pischke, 2008) departs from the specification of a linear model for the conditional mean of the potential outcomes. Then, any discrepancy between the observed outcome and that model is assumed to be guided by an statistical error whose properties are defined in terms of first and second order moments, as well as correlations across units. However, under randomization, we can derive the regression process from first principles and in particular as the result of the combination of SUTVA, the independence of random assignment and the potential outcomes and the strict use of pre-treatment variables, X. More precisely, let X be a $k \times 1$ vector of pre-treatment variables with expected value $E(X) = \mu_X$. Now, the potential outcomes might be a function of X, so that $Y_i(t, x_i)$. In the linear regression model, however, it is assumed that $Y_i(t) + X_i'\beta$. Under random assignment and pre-treatment X,

$$E\big[Y_i(1, X_i) - Y_i(0, X_i)|X, T\big] = E\big[Y_i(1) - Y_i(0)|X, T\big] = E\big[Y_i(1) - Y_i(0)\big] = \tau \quad (5.9)$$

As before, SUTVA is already implicit in this notation. For a given $X = x$, any variation in the potential outcomes is driven by the unit level error term

$$\varepsilon_i = (Y_i(0) - \alpha) + T_i \cdot \big(Y_i(1) - Y_i(0) - \tau\big) \quad (5.10)$$

where $\alpha = E(Y_i(0))$. From this it follows that

$$E(\varepsilon_i|T_i = 0, X) = E\big(Y_i(0) - \alpha|T_i = 0, X\big) = 0 \quad (5.11)$$

$$E(\varepsilon_i|T_i = 1, X) = E(Y_i(1) - \alpha - \tau|T_i = 1) = 0 \quad (5.12)$$

or $E(\varepsilon_i|T_i, X) = 0$. This assumption is allows us to specify a regression model for the observed outcome,

$$E(Y|X, T) = \alpha + \tau \cdot T + X'\beta \quad (5.13)$$

The parameters of this model can be estimated by Ordinary Least Squares,

$$(\tilde{\alpha}, \tilde{\beta}, \tilde{\tau}) = \underset{\alpha,\beta,\tau}{\operatorname{argmax}} \sum_{i=1}^{N} (Y_i - \alpha - X_i'\beta - \tau \cdot T_i)^2. \quad (5.14)$$

The conditional independence assumption $E(\varepsilon_i|T_i, X) = 0$ ensures that the Ordinary Least Squares (OLS) estimator of τ is consistent for the population parameter. Critically, consistency does not rest on the assumption that the linear specification $X'\beta$ is correctly specified (unlike it is customary when applying regression methods to observational studies). This thanks to a well known property of OLS (Greene, 2004). Under randomization of T, the OLS estimator of τ is unaffected by the values of X in repeated sampling -note, however, that the linearity assumption $Y_i(t, x_i) = Y_i(t) + X_i'\beta$ is essential for this result to hold. When X equals the empty set, it one can show that the OLS estimator equals the Neyman estimator

$$\tilde{\tau} = \hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \cdot T_i - \frac{1}{N_0}\sum_{i=1}^{N} Y_i \cdot (1 - T_i). \tag{5.15}$$

This endows the OLS estimator with a direct causal interpretation and it further implies that the OLS estimator in this case is also unbiased for the population parameter τ. Unsurprisingly, the sampling variance of the OLS estimator of τ in the absence of additional regressors X equals the estimator INSERTREF. In a classical randomized experiment setting, the inclusion of regressions does not matter to establish the consistency of the estimator. However it has relevance for the sampling variance of the estimator. Intuitively, if the regressors are relevant and provide better predictions for the levels of th outcome Y, the contribution of the residual variance to the overall sampling variance will be reduced. This can be better seen in by exploring the expression of the sampling variance of the OLS estimator of τ. It is not difficult to show that,

$$\tilde{\sigma}^2_{Y|W} = \frac{1}{N(N-1-K)} \frac{\sum_{i=1}^{N}(T_i - \bar{T})^2 \tilde{\varepsilon}_i^2}{(\bar{T}(1-\bar{T})^2)} \tag{5.16}$$

where $\tilde{\varepsilon}_i = Y_i - \tilde{\alpha} - \tilde{\tau} \cdot T_i - X_i'\tilde{\beta}$. The residuals $\tilde{\varepsilon}_i$ measures the discrepancy between a unit's observed level of outcome and the level of outcome predicted by the model, given T and X. They are, therefore, measuring lack of fitness. The contribution of the residuals to the estimated sampling variance will decrease if X is highly predictive of the potential outcomes. In turn this will increase the precision for the estimators of all the parameters, including $\hat{\tau}$.

With an estimator of the sampling variance it is simple to test hypothesis of the type $H_0 : \tau = \tau_0$ using the t-ratio

$$\nu' = \frac{\tilde{\tau}}{\sqrt{\tilde{\sigma}^2_{Y|W}}} \tag{5.17}$$

which has an asymptotic normal distribution. Therefore, give a value of the above ratio, ν^*, we would reject the null hypothesis of no statistical effect of the active treatment at

5% significance level if $|\nu'| > 1.96$.

1.6 Irregular assignment mechanisms.

The critical premise underpinning randomised experiments is that of unconfoundedness. Under this premise, the reception of an active or control treatment is independent of units' potential outcomes or characteristics. In practice, however, this premise is difficult to justify beyond the confines of pure experimental settings. In the social sciences, for example, researchers are often interested in understanding the consequences of different levels of educational attainment (Angrist and Lavy, 1999; Oreopoulos, 2006; Attanasio et al.; Card, 2001). Educational attainment is not randomly allocated. It depends on households' resources, parental background, local environment and, of course, each individual's own cognitive skills. Socio-economic status is correlated with families' investment in children's education. For example, Dahl and Lochner (2012) find that a $1,000 increase in income raises achievement in maths and reading test scores by 6 percent of a standard deviation in the short run. Similarly, Milligan and Stabile (2011) found that a child benefit program had significant positive effects on children's test scores (among other outcomes). A comprehensive review by Lochner and Monge-Naranjo (2012) presents further evidence that credit constraints have important implications for schooling (and other aspects of households' behaviour). In epidemiology, researchers often want to understand the effects of smoking and life-style on a population's overall health. Confounding when studying this kind of question arises because, again, people's decisions regarding specific health behaviours are likely to be driven by factors that are also correlated with health (such as educational attainment or socio-economic status; e.g. -Frisell et al., 2012; Kubicka et al., 2001; Stautz et al., 2016). Finally, even in controlled clinical environments, who receives a specific treatment or medical innovation in a trial might be driven by factors beyond randomization. Individuals assigned to the active treatment arm of an intervention might decide not to comply with the protocols or drop from the research. Conversely, individuals assigned to the control arm of an intervention might be able to force their way into the active treatment group if they intuit that the medical innovation might lead to health benefits. This might occur, for example, through sharing of information among participants or spillovers due to the benefits of the active treatment (e.g. vaccines).

Though confounding arises in a varied range of disciplines and its sources are varied, it is possible to bind the problem within the common framework of non-compliance with treatment assignment. When confounding is present in a study, the key to disentangling causal effects in data is to gain an understanding of the interaction between the assignment to treatment and the actual administered treatment. Specifically, empirical analysis of causal effects under confounding will rely on the characterisation of the problem at hand in terms of unconfounded *assignment* with confounded *treatment*. Within this framework,

the unconfounded assignment can be used to *predict* the theoretical treatment status of each unit or, more specifically, the discrepancy between the proportion of units treated given assignment status. Subsequently the predicted theoretical treatment assignment is used to identify the causal effect. In this sense, the unconfounded assignment acts as an *instrumental variable* in the analysis (Angrist, Imbens, and Rubin, 1996; Imbens and Angrist, 1994).

To illustrate this point, consider first the case of a randomised clinical experiment for a new cholesterol lowering medicine which happens to have an unintended and (at the point of delivery) unknown interaction with estrogen, a critical hormone for women. Imagine the new medicine creates an imbalance in estrogen leading to higher risk of anxiety among recipients. Suffering these effects, several women in the sample decide not to comply with the treatment. A number of men also fail to comply due to spurious reasons. At the time of evaluation the research team learns about these instances of non-compliance. The instrumental variable approach in this example will start by predicting the gap in treatment reception given assignment. Subsequently, any variation in outcomes between treated and untreated units will be proportionally attributed to those unit that complied with treatment only.

In the social sciences and epidemiology, the instrumental variable process is similar. However, it often involves working *backwards*: researchers typically observe a well defined treatment status and then an unconfounded assignment variable needs to be found to be able to identify the causal effect of treatment. A popular example is the estimation of the causal effect of retirement on health. There is a great concern that population ageing might lead to a large proportion of retirees in societies. This might put enormous pressure on the finance of social security systems around the world. An additional worry is that, at an individual level, retirement leads to drastic changes in people's lives: it removes a person from a daily activity, it leads to the perception of pensions which are often smaller than the working life salaries, and it multiplies the amount of free time. These changes might drive to changes in the nature of the decisions taken by people and end up affecting health. The sign of this effect, however, is more debatable, for retirement might often result in a relief from unhealthy environments, daily pressures and a reduction in anxiety levels. The critical aspect in this debate, however, is that retirement decisions might be driven by health itself, which disables the application of the methods in the previous section to estimate the treatment effect of interest (e.g. Disney et al., 2006). A peculiarity of retirement, however, is that it has been historically determined or incentivised by pension eligibility rules. Pension eligibility rules set a *normal* pension or early retirement age. These rules are exogenously determined by central governments and they are responsible for a significant number of retirement decisions across different nations every year (Gruber and Wise, 2002; Munnell et al., 2016). Researchers have seen a person's age in relation to governments normal pension age as determined by the accident

of birth, and thus as a form of unconfounded assignment. Whether a person then decides to retire or not upon reaching the pension age is likely to be determined by non-random factors. However, the proportion of theoretical retirements given the distribution of age in a sample is readily estimable. In an instrumental variable setting, variations in health could then be attributed to the proportion of the population retiring at the pension age.

As a final classic example of estimation under unconfounded assignment in the social science, consider Angrist's study of the effect military service on long term market outcomes (Angrist, 1998). Comparing labour market outcomes of a general population with the outcomes of Vietnam war veterans leads to a misleading consequences. Again, the problem is that individuals who serve in the military are likely to differ from the general population in non-trivial ways, for example, men with relatively few civilian opportunities might be likelier to serve in the military. Treatment (veteran status) in this case is not unconfounded. However, Angrist notes that eligibility to serve in the military during the Vietman war was partially determined by five draft lotteries. Eligible individuals were randomly allocated a number between 1 and 365 (identifying birth day). Weeks after the lottery, when the Defence Department knew its manpower needs, a ceiling number between 1-365 was set, and individuals with an allocated number below the ceiling were given priority for induction. This lottery, then, serves as an unconfounded assigment from which the likelihood of serving in the military can be predicted. In practice, Not everybody with a number under the ceiling served in Vietnam, and many individuals who might have not been eligible volunteered for service in order to gain favourable conditions. The important aspect here is that draft eligibility is not determined by traits such as people's potential for earnings. This allowed Angrist to attribute any variation in outcomes by veteran status to serving in the military.

The traditional textbook introduction to instrumental variable methods for confounded treatment and unconfounded assignment starts by introducing a regression model where the white noise process is, in fact, correlated with some of the regressors (e.g. Wooldridge, 2010; Angrist and Pischke, 2008). These regressors are called endogenous. Subsequently, an instrumental variable is defined which is correlated with the endogenous regressors but uncorrelated with the error term, and a theory of instrumental variables is devised leading to the so called two-stage least squares estimator. This approach, however, is developed on the basis of correlations which are difficult to interpret and, in any instance, empirically non-testable. Therefore, in the discussion that follows we take the modern approach introduced in Angrist et al., 1996, which makes clear statements regarding the conditions under which identification is possible and, more critically, which parameter is actually identifiable from data.

		Assignment 0	Assignment 1
Treatment	0	Never-takers Compliers	Never-takers Defiers
	1	Always-takers Defiers	Always-takers Compliers

Table 1.2: Subpopulations under confounded assignment

The potential outcomes framework in confounded observational studies.

As in the preceding discussion, data will be a sample of size N from a population. Although the potential outcomes are fixed at unit level, sampling creates a distribution of potential outcomes. We also maintain SUTVA throughout. We assume that the unconfounded assignment variable has already been identified (random assignment in an experiment; pension eligibility rule when studying retirement; lottery number in the Vietnam draft study). This variable will be denoted Z and takes on values 1 when a unit is assigned to the active treatment. Otherwise, Z equals 0. The actual treatment received might differ from the assigned treatment due to self-selection on the bases of personal characteristics. For instance, individuals with a frail health might be prone to retire earlier; individuals with relatively high lottery numbers might volunteer for military service in order to secure less exposed destination; participants in a clinical trial might not comply with treatment due to unforeseen side effects. However, and critically, Z is assumed to still determine treatment to a great extent. In other words, for each unit we define potential outcomes $T_i(Z_i)$.

We do not restrict compliance (both assigned and non-assigned units might comply or not). Therefore, four different types of units might co-exists in the population, described in Table 1.2. Units for which $Z = T$ are the compliers (with assignment). Some units might self-select into the active treatment regardless of assignment, so that $T = 1$ for any Z. These units are called always-takers. Similarly, some units might self-select into the control treatment group in any circumstance. These units are, consequently, never-takers. Finally, some units might be consistently rebellious, doing the opposite of what the assignment dictates. These units are termed defiers. What these definitions and Table 1.2 reveal is a fundamental problem of identification in observed data. Without further assumptions, data alone can not tell us which from which among the four populations a particular observation came from. To achieve that level of understanding, we will have to restrict the relative size of each subpopulation. Note that, in so far compliers and defiers exist in the population, it is accurate to define treatment as a function of the unconfounded assignment.

As in the case of randomized experiments, we can define potential outcomes for each unit. On this occasion, however, there is scope for these potential outcomes to depend on both treatment and assignment. Thus we define $Y_i(Z_i, W_i(Z))$. The dependence of the potential outcomes on Z will be a critical point in the discussion that follows.

Given that assignment is unconfounded, one might consider to apply the methods for randomized experiments in Section 1.5 to attribute variations in outputs to variations in assignment. In this respect, there are two parameters that could be estimated: the effect of assignment on treatment and the effect of assignment on outcomes,

$$ITT_T = E\Big[T_i(1) - T_i(0)\Big] \tag{6.1}$$

$$ITT_Y = E\Big[Y_i(1, T_i(1)) - T_i(0, T_i(0))\Big]. \tag{6.2}$$

These parameters are known as the Intention-to-Treat (ITT) estimands, given that they characterise the effect of unconfounded assignment -and, in the case of ITT_Y regardless of actual treatment received by each unit.

Let N_0, N_1 be the number of units assigned to the control and active treatment groups respectively. Under unconfoundedness, the distribution of Z is independent of $T(z)$ and $Y(z, T(z'))$ for every z, z', and we can estimate each of the above parameters

$$\hat{ITT}_T = \frac{1}{N_1}\sum_{i=1}^{N} T_i \cdot Z_i - \frac{1}{N_0}\sum_{i=1}^{N} T_i \cdot (1 - Z_i) \tag{6.3}$$

$$\hat{ITT}_Y = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \cdot Z_i - \frac{1}{N_0}\sum_{i=1}^{N} Y_i \cdot (1 - Z_i) \tag{6.4}$$

Under SUTVA and unconfounded assignment, these estimators are unbiased and consistent for the population counterparts. Under random sampling, the variance of the estimators conditional on the distributions of potential outcomes can be obtained as the sum of the variances of the assigned and un-assigned groups in the population, with estimators similar to those in equations (5.4) to (5.6). Confidence intervals and hypotheses test can then be derived as in the case of randomised experiments.

The shortcoming of ITT analyses is that, although they provide useful preliminary information, they are ultimately of limited use to understand the effect of the treatment under consideration. When studying the effect of retirement on health, understanding the effect of variations in pension eligibility rules on health are of natural policy concern, but the value of these estimates resides on the extent to which variation in these rules ultimately leads to variations in retirement patterns. When studying the effect of military service during the Vietnam war, understanding the behavioural effects of the lottery on service is of interest from the perspective of a policy planner. However, the parameter of ultimate consequences for the lives of people is the causal effect of serving in the military.

The instrumental variable method resolves the problem by arguing that any variation contained in the ITT_Y can be attributed solely to individuals who were affected by the assignment mechanism. The logic behind this arises from the observation that for the population of never-takers and always takers, the $ITT_Y = 0$ since $T(1) = T(0)$ for these individuals. This implies that any variation left can be attributed to either compliers or defiers. The method goes on to argue that defiers are unlikely in most settings and indeed one can rule out abnormal defying behaviours from any analysis. In a clinical trial, it would seem fanciful to comply with treatment by doing the opposite of what the research team suggests. As a result, if we attribute all the ITT_Y to the subpopulation of compliers in proportion to their presence in the sample, then the ratio,

$$LATE = \frac{ITT_Y}{ITT_T} \tag{6.5}$$

has a causal interpretation. This ratio is known as the Local Average Treatment Effect.

Underlying this logic there are a number of implicit assumptions, critical among which is the *exclusion restriction*. This restriction implies that unconfounded assignment matters only because of its effect on the treatment received. Specifically, the assignment variable does not have a direct effect on the outcome of interest. If this were the case, then it would be logically erroneous to attribute all the ITT_Y to the treatment variable -and indeed the ratio 6.5 would not be capture the LATE either.

We can develop this methodology formally. To this end, let us maintain SUTVA, and impose the following

Exclusion restriction For all z, z' and t, $Y(z,t) = Y(z',t)$.

The exclusion restriction implies that we can write $Y(Z, T(Z)) = Y(T(Z))$. Consider next the unit level treatment effect, $Y(1,T(1)) - Y(0,T(0))$. Under SUTVA and exclusion, this can be rewritten as:

$$\begin{aligned}
&Y(1,T(1)) - Y(0,T(0)) \\
&= \Big[Y(1)\cdot T(1) + Y(0)\cdot(1-T(1))\Big] - \Big[Y(1)\cdot T(0) + Y(0)\cdot(1-T(0))\Big] \\
&= \Big[\Big(Y(1)-Y(0)\Big)\Big(T(1)-T(0)\Big)\Big]
\end{aligned} \tag{6.6}$$

Averaging across the distribution of outcomes in the population, we have

$$\begin{aligned}
&E\Big[Y(1,T(1)) - Y(0,T(0))\Big] \\
&= E\Big[Y(1)-Y(0)|T(1)-T(0)=1\Big]\cdot P\Big(T(1)-T(0)=1\Big) \\
&- E\Big[Y(1)-Y(0)|T(1)-T(0)=-1\Big]\cdot P\Big(T(1)-T(0)=-1\Big)
\end{aligned} \tag{6.7}$$

The average unit-level treatment effect in the population equals the average unit level

treatment effect on the population of compliers, for whom $T(1) - T(0) = 1$ minus the average unit-level treatment effect for the population of defiers for whom $T(1) - T(0) = 1$. The difficulty comes because, as revealed by Table 1.2, we cannot identify from data alone which observations in a sample correspond to each subpopulations. However, if we could rule out the existence of defiers, we can then establish an estimable relationship between the population ITT_Y and the treatment effect for compliers. By definition, the subpopulation of defiers is an odd one, and it is often possible to succesfully argue that they do not exist in a specific publication. At the end of the day, however, this will have to be formalised in an non-refutable assumption: monotonicity.

Monotonicity $T(1) \geq T(0)$

This assumption implies that we can rewrite (6.7) as,

$$LATE = E\Big[Y(1) - Y(0)|T(1) - T(0) = 1\Big] = \frac{E\Big[Y(1, T(1)) - Y(0, T(0))\Big]}{E\Big[T(1) - T(0)\Big]} \tag{6.8}$$

Under the exclusion restriction, the numerator is ITT_Y. The denominator is ITT_T. Importantly, because of unconfounded assignment, both the numerator and denominator can be estimated consistently form the data using the techniques introduced for randomized experiments. Note that there is one final implicit assumption in operation here, namely that the denominator of the above ratio is not 0. In other words,

Nonzero average treatment effect of Z **on** T $E\Big[T(1) - T(0)\Big] \neq 0$

It is convenient to emphasise here that this results needs to hold only on average and so, for some units, assignment might indeed be of no substance. Thus, we are in no way restricting the existence of always- and never-takers.

We can estimate the LATE as

$$\hat{LATE} = \frac{\hat{ITT}_Y}{\hat{ITT}_T} \tag{6.9}$$

To obtain confidence intervals and construct test of hypotheses, we need an estimator of the sampling variance of this estimator. Using the delta method[1] we obtain the following

[1]Given the ratio of two random variables,

$$var\Big(\frac{R}{S}\Big) \approx \frac{1}{\mu_s^2} var(R) - 2\frac{\mu_R}{\mu_S^3} cov(R, S) + \frac{\mu_R^2}{\mu_S^4} var(S) \tag{6.10}$$

approximation to the sampling variance of the estimator,

$$var(\hat{LATE}) = \frac{1}{ITT_T^2} var(\hat{ITT}_Y) - 2\frac{ITT_Y}{ITT_T^3} cov(\hat{ITT}_Y, \hat{ITT}_T) + \frac{ITT_Y^2}{ITT_T^4} var(\hat{ITT}_T) \tag{6.11}$$

This approximation can be estimated by replacing population moments with sample analogues defined in the previous sections, as well as some estimator of the covariance of the estimated intention-to-treat parameters. The latter can be estimated via,

$$cov(\hat{ITT}_Y, \hat{ITT}_T) = \frac{1}{N_1 \cdot (N_1 - 1)} \sum_{i=1}^{N} Z_i \cdot (Y_i - \bar{Y}_1)(T_i - \bar{T}_1) \tag{6.12}$$

$$\bar{Y}_1 = \frac{1}{N_1} \sum_{i=1}^{N} Z_i \cdot Y_i \text{ and } \bar{Y}_0 = \frac{1}{N_0} \sum_{i=1}^{N} (1 - Z_i) \cdot Y_i \tag{6.13}$$

where N_t equals the number of units under treatment t.

Violations of the assumptions and weak instruments.

The identification strategy supporting the instrumental variable method hinges on the non-zero effect, monotonicity and exclusion restrictions. If any of these conditions is violated, then the ratio of intention-to-treat parameters does not provide accurate information about the effect of the active treatment on the group of compliers.

Consider, first, a violation of the exclusion restriction, holding all the remaining assumptions. In this case the assignment could affect the outcome of interest directly. Having ruled out the existence of defiers through monotonicity, this situation is characterised by always- or never-takers having a non-zero intention-to-treat effect on the outcome,

$$Y(1, t) - Y(0, t) \neq 0 \text{ for every } t \tag{6.14}$$

Then, the average unit effect of assignment in the population equals

$$\begin{aligned} &E(Y(1, D(1)) - Y(0, D(0))) \\ &= E(Y(1, D(1)) - Y(0, D(0)) | T(1) - T(0) = 1) P(T(1) - T(0) = 1) \\ &+ E(Y(1, D(1)) - Y(0, D(0)) | T(1) - T(0) = 0) P(T(1) - T(0) = 0) \end{aligned} \tag{6.15}$$

Noting that, in the absence of defiers, $E(T(1) - T(0)) = 1 \cdot P(T(1) - T(0) = 1) + 0 \cdot$

$P(T(1) - T(0) = 0)$, it follows that

$$\begin{aligned}
&\frac{E(Y(1, D(1)) - Y(0, D(0)))}{E(T(1) - T(0))} \\
&= E(Y(1, D(1)) - Y(0, D(0))|T(1) - T(0) = 1) \\
&+ E(Y(1, D(1)) - Y(0, D(0))|T(1) - T(0) = 0)\frac{P(T(1) - T(0) = 0)}{P(T(1) - T(0) = 1)}
\end{aligned} \tag{6.16}$$

Thus a failure of the exclusion restriction induces a bias proportional to, first, the odds of noncompliers in the population and, second, the effect of assignment on the population of non-compliers. The higher the influence of assignment on treatment (the stronger the instrument), the smaller this bias is, whereas the larger the effect assignment on the outcome, the larger the bias.

To illustrate how violations of exclusion can arise in practice, consider again a study of the effect of retirement on a random sample of single, elderly men's health, using pension eligibility as the instrumental variable. Suppose that in the labour market under study, no part-time employment or partial retirement were allowed. In that context, SUTVA appears to be a suitable assumption. Pension eligibility rules are exogenously determined by central governments and they are responsible for a significant number of retirement decisions across different nations every year (e.g. Gruber and Wise, 2002, Munnell et al., 2016). From this point of view, then, they seem to be a valid source of unconfounded assignments into retirement. The problem arises because pension eligibility rules tend to fix an eligibility age, typically around 60-65 years. These cut-off divides the population into older and younger groups at a time when the decline in people's health and cognitive functioning is accelerating. More precisely, those individuals who stay at work after reaching the eligibility age are subject to an average faster decline in cognitive functioning than other individuals. This creates a direct correlation between eligibility (a proxy for age) and cognitive functioning. That correlation violates the *exclusion* restriction. How severe is the bias affecting the estimator of the ratio of ITTs in this case? First, although pension eligibility rules do induce a number of retirements a year, the proportion of people retiring at the time is typically no more than 20%. In other words, the proportion of compliers in this case is restricted to just one fifth of the sample. Assuming that monotonicity holds, a direct implementation of the instrumental variable estimator will produce estimates that absorb a very large proportion of the direct effect of age on health. A potential strategy to reduce this bias is to try to focus on a narrow age band around the pension eligibility age, in order to mitigate the effect of age -however, this strategy is also likely to fail in practice (see Fé, 2018).

Violations of the monotonicity restriction are also troubling. Suppose that all other assumptions hold. Then, the ITT of Z on Y is zero for the never-takers, but there is a non-zero likelihood of defiers in the sample. In this instance, $E(T(1) - T(0)) =$

$P(T(1) - T(0) = 1) - P(T(1) - T(0) = -1) = \pi_c - \pi_d$. The ratio of intention-to-treat parameters now equals,

$$\begin{aligned}
&\frac{E(Y(1, D(1)) - Y(0, D(0)))}{E(T(1) - T(0))} \\
&= E(Y(1, D(1)) - Y(0, D(0))|T(1) - T(0) = 1) \\
&+ \lambda \cdot \Bigg[E(Y(1, D(1)) - Y(0, D(0))|T(1) - T(0) = 1) \\
&- E(Y(1, D(1)) - Y(0, D(0))|T(1) - T(0) = -1) \Bigg]
\end{aligned} \tag{6.17}$$

where,

$$\lambda = \frac{\Big[P(T(1) - T(0) = -1\Big]}{\Big[P(T(1) - T(0) = 1\Big] - \Big[P(T(1) - T(0) = -1\Big]} \tag{6.18}$$

The bias term now depends, first, on the difference in the treatment effect between compliers and non-compliers. Only if treatment affects both groups identically does the bias terms disappears. The second source of bias, λ, refers to the relative weight of the defiers in the population. If the proportion of defiers is large, or if the proportion of defiers and compliers is similar, then the bias term will be large, even if the treatment affects very similarly to compliers and defiers. Also, the stronger the proportion of compliers -the stronger the effect of the unconfounded assignment on treatment intake (or the stronger the instrumental variable)- the smaller will be bias term be.

Empirical research to day has paid considerably attention to violations in the exclusion restriction, however much less attention has been paid to violations of monotonicity, despite their potential to mislead the conclusions of empirical studies. Several sources, however, have emphasised the importance of these violation (de Chaisemartin, 2017; Fiorini and Stevens, 2014; Kitagawa; Angrist et al., 1996). Imbens and Angrist, 1994 94 provide specific examples of when monotonicity might be violated. Consider, for example, a group of candidates applying for a social security scheme, with two referees evaluating applications. Referee A accepts applicants with probability $P(0)$ while referee B (whose standards are more lenient) accepts applicants with probability $P(1) > P(0)$. Monotonicity here implies that any candidate approved by A should also be approved by B. Yet, suppose that A has a weakness for people with blue eyes, or pays particular attention to educational attainment (being more lenient towards higher achievers). If there are sufficient numbers of blue-eyed, highly educated individuals in the sample, the assumption of monotonicity is likely to be biased. de Chaisemartin, 2017 provides further examples of violations of the monotonicity assumption.

Implicit in all the discussion up to this point was the assumption that the unconfounded assignment variable was a strong instrument for the treatment effect, that is, Z conveys important information to explain the variation in the treatment indicator. The nonzero average treatment effect assumption is a necessary condition to ensure this is the case. However, non-zero average treatment effect is not sufficient. Specifically, we might find plausibly unconfounded instruments that might explain little about the variation in treatment. Such variables are termed weak instrumental variables, and have received considerable attention in recent years, particularly in the Econometrics literature (Nelson and Startz, 1990b; Nelson and Startz, 1990a; Bound et al., 1995; Staiger and Stock, 1997; **?**; Feir et al., 2016). The implications of weak instrumental variables are severe.

In the particular case of irregular assignment mechanisms weak instruments induce complications through at least two channels. In the extreme case when the instrument is spurious, so that $\text{ITT}_T = 0$, the LATE is not identifiable. However, in more common situations where the instrument carries some, but not much, information about the variation in treatment, the the most obvious consequence of a weak instrument is that the ITT_T will be small. Therefore, the LATE estimand will tend to be volatile over repeated sampling. This creates difficulties when testing hypotheses and estimating confidence intervals. Indeed, direct implementation of the methods discussed in this section will generally lead to unreliable tests of hypothesis. Numerous research articles have explored alternative inferential procedures for instrumental variable estimation under weak instruments (Davidson and MacKinnon, 2010; Davidson and MacKinnon, 2014; Staiger and Stock, 1997; Kleibergen, 2002; Andrews and Guggenberger, 2017; Mikusheva, 2010, Moreira, 2003 to mention but a few). Standard asymptotic critical values for the t-test in equation (5.17) based on the instrumental variable estimator lead to highly misleading inferences in the presence of weak instruments.

If weak instrumental variables lead to complications in inference, they are also problematic from a bias perspective. Bound et al., 1995 note that the use of weak instruments can lead to large inconsistencies in finite samples and, furthermore, the instrumental variable estimator is biased in the same direction than the intention to treat effect ITT_Y. Besides, weak instruments can considerably exacerbate the biases arising when some of the assumptions underpinning the instrumental variable estimator are violated. We have already pointed out that the violation of the exclusion or monotonicity assumptions lead to biases in the estimand of LATE. In both situations, however, the strength of the instrument was critical to determine the amount of bias. In the case of violations of the exclusion restriction, a weak instrument lead to over-emphasising the local effect of the active treatment among the group of non-compliers. In the case of violations of the monotonicity assumption, a weak instrument boosted exacerbated the bias due to the contribution of the effect of the subpopulation of defiers to the estimand of the LATE. Thus, a weak instrumental variable can considerably exacerbate any bias due to violations

of the exclusion or monotonicity assumptions.

One of the most debated instrumental variables in the empirical literature is found in Angrist and Krueger (1991). These authors were interested in understanding the effect of schooling on educational attainment and earnings later in life. As these authors note, many countries around the world have established compulsory education, typically until the mid/late teens, but the contribution of education to later life outcomes is not entirely well understood. This is also a difficult question to investigate with observational studies because schooling is likely to be driven by innate personal and environmental factors that also determine personal income and educational attainment. Angrist and Krueger (1991), however, note that compulsory schooling laws compel students born in certain months to attend school longer than other months. Specifically, in countries such as U.S. and Spain, children start to attend school in the late summer (typically September) of the year they reach a specific age. This implies that a child born in January starts school at an older age than a child born in December. It also implies, at least in theory, that if the compulsory schooling law is stopping some children from leaving school, those drop-outs born early in the year will have completed less time in school than drop outs born later in the year.

Angrist and Krueger (1991) present abundant evidence to support this assumption. Specifically, looking at Census data from the U.S. they discover a seasonal patter suggesting that students born early in the year complete, on average, about 0.1 years of education less. Arguing that the date of birth is unlikely to be correlated with personal attributes, AK conclude that quarter of birth is an unconfounded predictor of educational attainment, and thus a valid instrument to explore the effect of schooling on life outcomes.

In a follow up study, however, Bound et al., 1995 note that the results in Angrist and Krueger, 1991 are fairly sensitive to the inclusion of covariates and the selection of instrumental variables in Angrist and Krueger's model. Furthermore, the instrumental variable estimator of the causal effect tended towards the estimate produced by a least squares regression. This suggested that perhaps quarter of birth are, individually, poor predictors of schooling, but when combined a spurious ability to explain schooling might arise. Indeed, Bound et al., 1995 re-run Angrist and Krueger's analysis using a randomly allocated quarter of birth as the instrument. This latter analysis closely replicated the results obtained from the least squares regression and, therefore, also the instrumental variable analysis. Using additional theoretical arguments, Bound et al. (1995) and late Staiger and Stock, 1997 conclude that quarter-of-birth is indeed a weak instrumental variable for schooling.

Other irregular assignments in practice.

As noted above, estimation of (local) causal effects via instrumental variables is a retrospective method. The research question comes first and the researcher has to subsequently search for a suitable instrumental variable. The latter, however is not a trivial process and

requires a good understanding of the environmental and institutional context surrounding the main research question. Empirical researchers, however, have been successful at identifying a large catalogue of valid instrumental variables. These instruments exploit a variety of phenomena and institutional peculiarities, including lotteries, weather events, natural disasters, temporal variation in policies or industrial configurations, historical events, eligibility rules, quarter of birth and cut-offs in assignment rules. The quality of instruments in terms of their predictive power and their ability to satisfy the unconfoundedness, exclusion and monotonicity restrictions varies vastly, with new research often debunking previously well established instruments.

One of the most favourable settings to find strong instrumental variables arises when underlying the treatment of interest there is a process of randomization. A good example of this is the *National Job Corps Study* . Job Corps is no-cost educational and vocational training program administered by the United States Department of Labor. The aim of the program is to help 16 to 24 year old individuals from disadvantaged backgrounds to become more employable or, more specifically, to *improve the quality and satisfaction of their lives through vocational and academic training.* The program was inaugurated in 1964. At the time youth unemployment doubled the unemployment registered among the rest of society. The initial remit of the scheme was circumscribed to the US Federal National Parks, National Forests, and other Federal Lands, however it has now expanded to cover most professional fields.

The program serves more than 60,000 new participants every year at an estimated cost in 2014 of about 1.7 billion US dollars. Give the ambitious goals of the program and its cost (it constitutes more than 60% of all funds spent by the US Department of Labor), assessing its effectiveness has been a main concern of academics and policy-makers alike. An evaluation of the program by comparing the outcomes of participants and non-participants will lead to misleading conclusions because, to begin with, participants in the Job Corps will differ from non-participants in socioeconomic background.

In 1993, the Department of Labor introduced the National Job Corps Study, a nationally representative experimental evaluation. Between 1994 and 1996, over 80,000 eligible applicants were allocated to an active treatment group (effectively joining Job Corps) or a control group, being effectively excluded from the program (in practice, applicants in the control group were embargoed from the program for three years). Following randomization, interviews were planned after 52, 130 and 208 weeks.

The design of the program suggests that the causal effect of training could be evaluated as a randomized experiment. However, there were a number of difficulties which precluded the use of the methods described in section 5. The most critical difficulty for the present discussion was that compliance with assignment was imperfect, with only 68% of eligible applicants in the active treatment group enrolling and participating in Job Corps for at least one week.

In the absence of further complications (such as missing response data) evaluating the effect of Job Corps by focusing on the variation of outcomes by assignment (that is an intention-to-treat study) will fail to capture the policy parameter of interest -if Job Corps is indeed successful, the intention-to-treat parameter will under-estimate the impact of the program by pooling non-compliers alongside individuals in the control group. However, random allocation to treatment in the Job Corps Study opens the door for an instrumental variable analysis. The randomization scheme could explain 60% of the intake in the assigned group and allocation to treatment groups was uncorrelated with applicant's observable or unobservable traits and is unlikely to have any direct effect of candidate's future outcomes other than through its effect on intake. In other words, we can postulate the random assignment as an unconfounded, strong instrumental variable for enrolment into Job Corps. Then, the instrumental variable methods discussed in this chapter can be used to evaluate the program. Various evaluation of the program can be found in LaLonde, 1986, Heckman et al., 1997, Lee (2009), Schochet et al., 2008 or Flores and Flores-Lagunes, 2013, to mention but a few.

In the opposite spectrum of studies where randomization can be used as an instrumental variable, we find studies of much more diffuse causal effects, where considerable creativity are required together with significant efforts to compile a suitable data set. Two recent studies illustrate this point well.

In a highly novel study, Borowiecki (2017) addresses the question of whether or not emotions and mood have a causal effect on creativity. Addressing this question is complex. Not only can creativity and emotions feed into each other, but measuring them is complicated. Individuals' self-assessments of their own emotions are often unreliable, subject to measurement error, often not being comparable across individuals (who have different levels of tolerance for different emotional states). Creativity is an equally elusive construct, however on the basis of previous research Borowiecki (2017) concludes that 'eminent' creativity (the Big-C) is potential the object of primary interest, as only outstanding creativity that has a long-lasting legacy and the potential to change the course of different fields in the arts and the sciences. In view of these limitations, Borowiecki (2017) construct a unique data set spanning the lives of three well known composers: Beethoven, Mozart and Liszt. To measure their emotional states, Borowiecki (2017) uses linguistic software to constructs lifetime well-being indices using around 1,400 letters written by these composers. The focus is particularly on negative emotions as Borowiecki (2017) find preceding literature establishing a correlations between creativity and negative emotions. Creativity is then measures as measured as the number of important, quality-adjusted compositions written by each composer in a given year. The instrumental variable for emotions is the unexpected death of a composer's family member. Unexpected deaths have been used as an instrumental variable in other areas of research, most notably when studying the effects that the death of top academics have on their networks of collabora-

tors and their sub-fields of research Azoulay et al., 2010, Azoulay et al. (2015). With this identification strategy, Borowiecki (2017)finds that the number of works written is causally attributable to an increase in negative emotions, in particular, among anger, anxiety, and sadness, is the latter which appears to be the main feeling that drives creativity.

1.7 Partial Identification Approach.

The defining feature of causal analysis is the problem of selection: we can only observe one potential outcome per unit. As a result, data alone cannot answer causal questions and we need to imagine what the level of the unobserved counterfactual outcome might be, at least on average, in the population. In this respect, numerous modelling techniques have been put forward to complete the gap left by unobserved counterfactuals. These include Difference-in-Differences (classic examples are Ashenfelter and Card, 1985; Card and Krueger, 1994; Duflo, 2001, Poterba et al., 1995, Blundell et al., 2004 Athey and Imbens, 2006), control function analyses (e.g. Heckman (1976); Ahn and Powell (1993); Andrews and Schafgans, 1998;) and the now pervasive Regression Discontinuity Design (Thistlethwaite and Campbell, 1960; Hahn, 1998; Lee and Card, 2008; Card et al., 2015; Imbens and Kalyanaraman, 2012; Cattaneo et al., 2015; Frandsen et al., 2012; Calonico et al., 2014; excellent introductory surveys are Lee and Lemieux, 2010 and Imbens and Lemieux, 2008). All these methods can be framed as variations on the topic of instrumental variables. For example, a basic Fuzzy Regression Discontinuity Design is, from the strict and narrow point of view of implementation, a ratio of intention to treat parameters restricted to a carefully selected neighbourhood around a fixed policy cut-off point. The important aspect determining the suitability of one method over another in a specific empirical study is the assumptions underlying the method regarding the omitted potential outcomes and the assignment mechanism. These assumptions are far from naive technicalities required by suitable methods; rather are the driver of the conclusions of causal analysis.

Assumptions are essential to answer causal questions. However, how certain can we, in general, be about the appropriateness of a set of assumptions? Said differently, how certain can we be about the levels of unobserved counterfactuals? How certain can we be about the workings of unobservable assignment mechanisms and the reasons why some units received the active treatment and some other units did not? As with matters of religion, in science the credibility of assumptions tends to be a subjective matter. In scientific matters, uncertainty is a natural state and characterising this uncertainty should be a primary goal of science. This point has been recurrently emphasised in a body of research which can be traced back to Frechet (1951), Marschak and Andrews (1944) but which has gained substantial momentum over the last four decades thanks to the seminal contributions of Charles F. Manski. This body of research has emphasised the idea of

Partial Identification:

'*[...] first ask what can be learned from data using knowledge of the sampling process alone, then ask what more can be learned when data are combined with weak but widely credible distributional assumptions, and finally ask what further can be learned when the data are combined with stronger, less credible assumptions* (Manski (2007))'.

In this context, identification is partial because weak credible assumptions lead to statistical methods which result in ranges of values binding the parameter of interest, as opposed to point estimates exactly quantifying that parameter.

Two main motivations underlie the Partial Identification approach. First, there is Manski's Law of Decreasing Credibility (Manski (2003)). This law suggests that the more forcefully assumptions need to be imposed (the stronger these assumptions are), the least credible inferences will seem and the more embattled a causal statistical analysis will be. On the contrary, the view of the Partial Identification approach is that weak assumptions help to establish a '*domain of consensus*' and bind researchers' disagreements Manski (2018). More fundamentally, however, the Partial identification literature sees sciences as a 'social enterprise in that:

'[...] *disregard of uncertainty when reporting research findings may harm formation of public policy. If policy makers incorrectly believe that existing analysis provides an accurate description of history and accurate predictions of policy outcomes, they will not recognize the potential value of new research aiming to improve knowledge. Nor will they appreciate the potential usefulness of decision strategies that may help society cope with uncertainty and learn, including diversification and information acquisition.* (Manski, 2018)'

The social dimension of partial identification, combined with its emphasis on quantifying the natural uncertainty surrounding science and the sophistication of the underpinning statistical techniques make the approach one of the most important contributions to the field of statistics in recent years. This last section presents some of the best understood ideas on Partial identification of population average treatment effects. The framework has, however, been widely extended to a variety of settings (see, among others, Honor and Lleras-Muney (2006); Kreider and Pepper (2007); Horowitz and Manski (2000); Blundell et al.; Manski and Pepper (2017); Manski and Pepper (2013); Ciliberto and Tamer (2009); Kreider et al. (2012); Galichon and Henry (2011); Okumura and Usui (2014); Chernozhukov et al. (2013), Beresteanu and Molinari (2008)). For excellent surveys see Tamer (2010) and Ho and Rosen (2017). In this survey, we will focus on applications to the most popular methods for the partial identification of causal effects starting with Manski (1990). Inference in the partial identification setting is complex and standard

procedures, including the bootstrap, cannot be applied directly to draw inferences about the underlying parameters and identification regions. This is an area of active research, but we do not cover it in this review. The interested reader can refer to Imbens and Manski (2004), Chernozhukov et al. (2013), Bugni (2010), Andrews and Soares (2010), Chernozhukov et al. (2007) and references therein.

The starting point of the partial identification approach for causal analyses is the problem of selection: a unit's potential outcomes $Y(0)$ and $Y(1)$ cannot be observed simultaneously. The implications and consequences of this can be summarised in the following equality,

$$\begin{aligned} E[Y(t)] &= E(Y(t)|T=t)P(T=t) + E(Y(t)|T\neq t)P(T\neq t) \\ &= E(Y|T=t)P(Z=z) + E(Y(z)|Z\neq z)P(Z\neq z). \end{aligned} \tag{7.19}$$

for $t \in \{0,1\}$. Data are not available to estimate the conditional means $E(Y(t)|T \neq t)$ and so it is not possible to point identify the effect of T, namely $E\Big[Y(1)-Y(0)\Big]$, unless one is willing to assume that treatment is unconfounded (as would occur in a classic randomized experiment). In that case, it would hold that treatment assignment is uncorrelated or independent of unit's personal traits, so that $E(Y(t)|T \neq t) = E(Y(t)|T = t)$ and so a difference in means among treated and non-treated units will constitute an unbiased and consistent estimate of the causal effect of the active treatment on the outcome of interest.

As the discussion in the preceding section made clear, however, in observational studies the reception of treatments T is rarely unconfounded. In principle, we could try to pursue an instrumental variable strategy. However, even in the case of a perfectly designed instrumental variable environment, with strong, relevant instruments that only affect the outcome through their effect on the uptake of the treatment, the estimation of causal effects in irregular assignment mechanisms reveals a Local Average Causal Effect. This is often an interesting parameter, but it's scope is circumscribed to the subpopulation of compliers. The latter subpopulation might be poorly representative of the whole population. For example, studies about the effect of retirement on consumption, health and cognition abound (Mazzonna and Peracchi, 2012; Bonsang et al., 2012;Banks et al., 1998). Most of these studies use pension eligibility rules as an instrumental variable for retirement decisions. Leaving aside considerations regarding the validity of this instrument (see Fé, 2018), the population of compliers in this setting tends to be between 7-20% depending on the country of study and the data set. Policy-makers will find these studies of interest, because they have the ability to set and modify pension eligibility rules and the effect of these rules is clearly important. However, it would be daring to design a welfare policy for the whole population based on the responses and experiences of just the 7-20% of compliers in the population. In summary, the average treatment effect (as opposed to the Local Average Treatment Effect) it still the critical parameter $E\Big[Y(1)-Y(0)\Big]$ in

causal studies.

One might naturally wonder what can be learned about the effect of T without relying on that assumption or indeed '*any*' other substantive assumption. This question was first studied by Manski (1990) who noted that, if the range of variation of Y is bounded within a finite interval $[a, b] \subset \mathbb{R}$, then we can bind the effect of interest within the following identification region:

$$\begin{aligned} \text{ATE} &\in \left[\underline{\theta}_{\text{nab}}, \overline{\theta}_{\text{nab}}\right] \\ \underline{\theta}_{\text{nab}} &= \Big[E(Y|T=1) - b\Big] \cdot P(T=1) - \Big[E(Y|T=0) - a\Big] \cdot P(T=0) \\ \overline{\theta}_{\text{nab}} &= \Big[E(Y|T=1) - a\Big] \cdot P(T=1) + \Big[b - E(Y|T=0)\Big] \cdot P(T=0) \end{aligned} \tag{7.20}$$

The bounds $\overline{\theta}_{\text{nab}}$ and $\underline{\theta}_{\text{nab}}$ are often referred to as the No Assumption Bounds (NAB), a terminology somehow misleading, given the underlying assumption of bounded potential outcomes. Nonetheless, we adopt this term here for convenience. The bounds in the preceding equations can be estimated using sample analogues to replace conditional expectations in (7.20). As can be inferred from the definitions of $\underline{\theta}_{\text{nab}}$ and $\overline{\theta}_{\text{nab}}$, 0 is always included in the identification region and, as a result, not even the sign of the effect of T identified. Therefore, despite of their undeniable credibility, these bounds are of little value from a policy perspective (an observation already made in Manski, 1990).

The NAB are highly credible but of not very informative. This has led to substantial research to find alternative sets of assumptions that, being credible, result in more informative bounds. In what follows, we consider three of these assumptions: Monotone Treatment Response, Monotone Treatment Selection and Monotone Instrumental Variable. These constitute the better understood techniques in the Partial Identification literature. They also present a hierarchy in terms of credibility and identification power.

In certain situations, we will be able to establish a ranking among units' potential outcomes. Specifically, it will often be reasonable to assume that, for example, $Y(1) \leq Y(0)$, implying that treatment cannot have a long term positive effect on outcomes (in other settings the converse might be more suitable, namely $Y(1) \geq Y(0)$). For example suppose that T corresponds to the death of close relative, as in Borowiecki (2017) study, or a natural disaster, such as a drought (e.g.Shah and Steinberg (2017)). This assumption is known as the Monotone Treatment Response (MTR) assumption and was introduced in the seminal paper by Manski (1997). MTR constitutes a first step towards obtaining narrower partial identification regions than the no-assumption bounds. Intuitively, the ranking is restricting the sign of the unit level causal effect and, therefore, what MTR does is to set an upper bound on the average treatment effect at 0 (in the case when $Y(1) \geq Y(0)$, MTR imposes a lower bound at 0). This can be seen formally by noting

that under MTR,

$$\begin{aligned} E(Y(0)) \in & \Big[E(Y|T=0)\cdot P(T=0) + E(Y|T=1)\cdot P(T=1), \\ & E(Y|T=0)\cdot P(T=0) + b\cdot P(T=1)\Big] \\ E(Y(1)) \in & \Big[a\cdot P(T=0) + E(Y|T=1)\cdot P(T=1), \\ & E(Y|T=0)\cdot P(T=0) + E(Y|T=1)\cdot P(T=1)\Big] \end{aligned}$$

and so,

$$ATE \in \Big[\underline{\theta}_{\text{nab}}, 0\Big] \tag{7.21}$$

The power of MTR comes from restricting the sign of the ITT to the non-negative real line. However, MTR does not provide any information regarding the magnitude of the lower bound of the identification region (which equals to the bound obtained under NAB). Therefore, the MTR remains somehow uninformative about the average treatment effect. An additional drawback of MTR is that it must hold exactly across all units. This might be too restrictive.

MTR draws its identification power from substantive assumptions that restrict the behaviour of the potential outcomes. Alternatively, one can make assumptions about the underlying assignment mechanism and, in particular, what might drive the variation in treatment across units. One such assumption is the Monotone Treatment Selection (MTS) assumption in Manski and Pepper (2000). MTS does not fully rule out selection into treatment and indeed it allows treated individuals to differ in non-trivial ways from those who have not been subject to treatment, provided that, on average, we can establish a monotone relationship explaining selection into treatment. Formally, MTS implies that for either $t \in \{0,1\}$, $E(Y(t)|T=0) \geq E(y(t)|T=1)$. Once again, it is convenient to emphasise that MTS is a substantive assumption and, indeed, a researcher might find the above arguments unconvincing. In that case, a different set of assumptions could be put forward instead.

When we impose the MTS assumption, we find that

$$\begin{aligned} E(Y(0)) &\in [E(Y|T=0)P(T=0) + aP(T=1), E(Y|T=0)] \\ E(Y(1)) &\in [E(Y|T=1), E(Y|T=1)P(T=1) + bP(Z=0)] \end{aligned}$$

and so,

$$\begin{aligned} &\text{ATE} \in \Big[\underline{\theta}_{\text{mts}}, \overline{\theta}_{\text{nab}}\Big] \\ &\underline{\theta}_{\text{mts}} = E(Y|T=1) - E(Y|T=0). \end{aligned} \tag{7.22}$$

MTS leaves the upper bound under NAB intact, but it considerably increases the lower bound, which is set equal to the difference in mean outcomes across assignment groups. The latter is the estimate of the ATE in a randomized experiment or observational study with unconfounded treatment. In other words, when the assumption of unconfoundedness is not credible, but one can reasonably establish the sign of the assignment mechanism, then the estimator of the treatment effect under unconfoundedness can only be interpreted as an upper/lower bound (depending on the assumed sign of the assignment mechanism) .

An interesting observation made by Manski and Pepper (2000) is that MTS and MTR can be combined to obtain an even more informative identification region yet. Specifically, consider the non-negative version of MTS, so that $E(Y(t)|T=0) \leq E(y(t)|T=1)$. Then, it is not difficult to show that

$$E(Y|T=0) \leq E(Y(0)) \leq E(Y|T=0)P(T=0) + E(Y(0)|T=1)P(T=1)$$
$$E(Y|T=1)P(T=1) + E(Y(1)|T=0)P(T=0) \leq E(Y(1)) \leq E(Y|T=1)$$

We can bring a non-negative version of MTR on board, $Y(1) \geq Y(0)$, to further narrow the above bounds. In particular, note that

$$E(Y(0)|T=1) \leq E(Y|T=1) \tag{7.23}$$

and

$$E(Y(1)|T=0) \geq E(Y|T=0) \tag{7.24}$$

So we find that MTR and MTS, when combined, imply,

$$E(Y|T=0) \leq E(Y(0)) \leq E(Y|T=0)P(T=0) + E(Y|T=1)P(T=1)$$
$$E(Y|T=1)P(T=1) + E(Y|T=0)P(T=0) \leq E(Y(1)) \leq E(Y|T=1)$$

and it follows that

$$0 \leq ATE \leq E(Y|T=1) - E(Y|T=0) \tag{7.25}$$

As noted by Manski and Pepper (2000), the combination of MTS and MTR leads to a testable implication, namely $E(Y|T=1) \geq E(Y|T=0)$. This condition can be tested in practice using a standard one-sided t-test of the null hypothesis $H_0 : E(Y|T=1) = E(Y|T=0)$ and alternative hypothesis $H_1 : E(Y|T=1) \geq E(Y|T=0)$. If the null hypothesis is rejected we do not conclude that MTR and MTS are true, however. Instead we take this a evidence that the combination of these two assumptions *might* be supported by the data.

MTS is a particular instance of a Monotone Instrumental Variable. The instrumental variables considered in the preceding section had to be unconfounded in the sense that

they affect outcomes only through their effect on the intake of treatment but, otherwise, they are independent of a unit's potential outcomes. This assumption can be contentious in practice and indeed empirical studies tends to go considerable lengths to try to justify this kind of assumption. Manski and Pepper (2000) note that, whereas it is often possible to argue that candidate instruments violate the strict unconfoundedness assumption, it is also possible to find numerous situations when the relationship between the candidate instrument and the potential outcomes is monotonic. In that case, one can use this 'invalid' instrumental variable to obtain credible information about the underlying treatment effect. Specifically Manski and Pepper (2000) introduce the concept of a Monotone Instrumental Variable (MIV).

Suppose there is a variable W such that, for every z, ζ and $w_1, w_2 \in (\mathcal{W} \times \mathcal{W})$, the following assumption might be reasonable

$$E(Y(t)|T = \zeta, W = w_1) \leq E(Y(t)|T = \zeta, W = w_2) \text{ whenever } w_1 \leq w_2. \tag{7.26}$$

The variable W is known as a Monotone Instrumental Variable: it restricts the relationship between mean responses across the treatment and control groups, but it fall short of imposing a this relationship constant (as a standard Instrumental Variable does; e.g. Greene, 2004).

Let $LB(t|w)$, $UB(t|w)$ denote the bounds for $E(Y(t)|W = w)$ under NAB, MTR or MTS. For $w_1 \leq w \leq w_2$, Manski and Pepper (2000) show that $E(Y(t)|W = w_1) \leq E(Y(t)|W = v) \leq E(Y(t)|W = w_2)$ and therefore,

$$\max_{w_1 \leq w} \{LB(z|w_1)\} \leq E(Y(t)|W = w) \leq \min_{w_2 \geq w} \{UB(t|w_2)\}. \tag{7.27}$$

Given that (7.26) implies,

$$E(Y(t)) = \sum_{w \in \mathcal{W}} E(Y(t)|W = w)P(W = w),$$

we have

$$\begin{aligned} &\sum_{w \in \mathcal{W}} P(W = w) \left\{ \max_{w_1 \leq w} \{LB(t|w_1)\} \right\} \\ &\leq E(Y(t)) \\ &\leq \sum_{w \in \mathcal{W}} P(W = w) \left\{ \min_{w_2 \geq w} \{UB(z|w_2)\} \right\} \end{aligned} \tag{7.28}$$

At any given w, equation (7.27) replaces the conditional lower (upper) bounds under NAB, MTS or MTR for a larger (lower) value in those areas in the support of the MIV where the alleged monotonic relationship between Y and W is not empirically satisfied.

This results is narrower identification regions for $Y(z)$ and, consequently, for the average treatment effect.

Bibliography

Ahn, H. and **Powell, J.L.** (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1): 3 – 29

Andrews, D.W.K. and **Schafgans, M.M.A.** (1998). Semiparametric estimation of the intercept of a sample selection model. *The Review of Economic Studies*, 65(3): 497–517

Andrews, D.W.K. and **Soares, G.** (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1): 119–157

Andrews, D.W. and **Guggenberger, P.** (2017). Asymptotic size of kleibergens lm and conditional lr tests for moment condition models. *Econometric Theory*, 33(5): 10461080. doi:10.1017/S0266466616000347

Angrist, J.D. and **Pischke, J.** (2008). *Mostly Harmless Econometrics*. Princeton University Press

Angrist, J.D. (1998). Estimating the labor market impact on voluntary military service using social security data on military applicants. *Econometrica*, 66: 249–288

Angrist, J.D., **Imbens, G.W.**, and **Rubin, D.** (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91: 444–455

Angrist, J.D. and **Krueger, A.B.** (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4): 979–1014

Angrist, J.D. and **Lavy, V.** (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement*. *The Quarterly Journal of Economics*, 114(2): 533–575

Ashenfelter, O. and **Card, D.** (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4): 648–60

Athey, S. and **Imbens, G.W.** (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2): 431–497

Attanasio, O., **Meghir, C.**, and **Santiago, A.** (). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate Progresa. *Review of Economic Studies*

Azoulay, P., **Fons-Rosen, C.**, and **Wang, J.** (2015). Does Science Advance One Funeral at a Time? *NBER Working Paper*, 21788

Azoulay, P., **Graff Zivin, J.S.**, and **Wang, J.** (2010). Superstar Extinction. *The Quarterly Journal of Economics*, 125(2): 549–589

Balke, A. and **Pearl, J.** (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439): 1171–1176

Banks, J., **Blundell, R.**, and **Tanner, S.** (1998). Is there a retirement-saving puzzle? *American Economic Review.*, 88: 769–788

Battistin, E., **Brugiavini, A.**, **Rettore, E.**, and **Weber, G.** (2009). The retirement consumption puzzle: Evidence from a Regression Discontinuity approach. *American Economic Review*, 99: 2209–2226

Beresteanu, A. and **Molinari, F.** (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4): 763–814

Bickel, P.J., **Hammel, E.A.**, and **O'Connell, J.W.** (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175): 398–404

Blundell, R., **Dias, M.C.**, **Meghir, C.**, and **van Reenen, J.** (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, 2(4): 569–606

Blundell, R., **Gosling, A.**, **Ichimura, H.**, and **Meghir, C.** (). Changes in the distribution of male and female wages accounting for employment composition usingbounds. *Econometrica*, 75(2): 323–363. doi:10.1111/j.1468-0262.2006.00750.x

Bonsang, E., **Adam, S.**, and **Perelman, S.** (2012). Does retirement affect cognitive functioning? *Journal of Health Economics*, 31: 490–501

Borowiecki, K.J. (2017). How are you, my dearest mozart? well-being and creativity of three famous composers based on their letters. *The Review of Economics and Statistics*, 99(4): 591–605. doi:10.1162/REST_a_00616

Bound, J., **Jaeger, D.**, and **Baker, R.** (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90: 443–450

Bugni, F.A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2): 735–753

Calonico, S., **Cattaneo, M.D.**, and **Titiunik, R.** (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6): 2295–2326

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5): 1127–1160

Card, D. and **Krueger, A.** (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4): 772–93

Card, D., **Lee, D.S.**, **Pei, Z.**, and **Weber, A.** (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6): 2453–2483

Cattaneo, M., **Frandsen, B.**, and **Titiunik, R.** (2015). Randomization inference in the Regression Discontinuity Design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, 3: 1–24

Charles, K. (2004). Is retirement depressing? Labor force inactivity and psychological well-being in later life. *Research in Labor Economics*, 23: 269–299

Chernozhukov, V., **Hong, H.**, and **Tamer, E.** (2007). Estimation and confidence regions for parameter sets in econometric models1. *Econometrica*, 75(5): 1243–1284

Chernozhukov, V., **Lee, S.**, and **Rosen, A.M.** (2013). Intersection bounds: estimation and inference. *Econometrica*, 81(2): 667–737

Ciliberto, F. and **Tamer, E.** (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6): 1791–1828

Coe, N. and **Zamarro, G.** (2008). Retirement effects on health in Europe. Technical report, RAND Labor and Population Working Paper W-588

Cook, T.D. (2018). Twenty-six assumptions that have to be met if single random assignment experiments are to warrant gold standard status: A commentary on deaton and cartwright. *Social Science and Medicine*, 210: 37 – 40. ISSN 0277-9536. doi: https://doi.org/10.1016/j.socscimed.2018.04.031. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue

Cox, D. (1958). *Planning of experiments*. Wiley-Interscience. Wiley

Cox, D. (2009). Randomization in the design of experiments. *International Statistical Review*, 77(3): 415–429

Dahl, G.B. and **Lochner, L.** (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *American Economic Review*, 102(5): 1927–56. doi:10.1257/aer.102.5.1927

Davidson, R. and **MacKinnon, J.** (2010). Wild bootstrap tests for IV regression. *Journal of Business and Economic Statistics*, 28(1): 128–144

Davidson, R. and **MacKinnon, J.G.** (2014). Bootstrap confidence sets with weak instruments. *Econometric Reviews*, 33(5-6): 651–675

de Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2): 367–396

Deaton, A. and **Cartwright, N.** (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, 210: 2 – 21. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue

Diamond, J. and **Robinson, J.** (2011). *Natural Experiments of History*. Harvard University Press

Disney, R., **Emmerson, C.**, and **Wakefield, M.** (2006). Ill health and retirement in Britain: A panel data-based analysis. *Journal of Health Economics.*, 25: 621–649

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *The American Economic Review*, 91(4): 795–813

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Strategies for Social Inquiry. Cambridge University Press. doi:10.1017/CBO9781139084444

Fé, E. (2018). Pension eligibility rules and the local causal effect of retirement on cognitive functioning. Technical report, Social Science Research Network. Https://ssrn.com/abstract=2993152

Feir, D., **Lemieux, T.**, and **Marmer, V.** (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 34(2): 185–196

Fiorini, M. and **Stevens, K.** (2014). Assessing the monotonicity assumption in iv and fuzzy rd designs. Working Papers 2014-13, University of Sydney, School of Economics

Fisher, R. (1935). *The Design of Experiments*. Oliver Boyd

Flores, C.A. and **Flores-Lagunes, A.** (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business and Economic Statistics*, 31(4): 534–545

Frandsen, B., **Frolich, M.**, and **Melly, B.** (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168: 382–395

Frechet, M. (1951). Sur les tableaux de correlation dont les marges sont donnes. *Ann. Univ. Lyon A*, 3: 53–77

Frisell, T., **Pawitan, Y.**, and **Långström, N.** (2012). Is the association between general cognitive ability and violent crime caused by family-level confounders? *PloS one*, 7: e41783. ISSN 1932-6203

Galichon, A. and **Henry, M.** (2011). Set Identification in Models with Multiple Equilibria. *The Review of Economic Studies*, 78(4): 1264–1298

Greene, W. (2004). *Econometric Analysis*. Prentice Hall

Gruber, J. and **Wise, D.** (2002). *Social Security and retirement around the World*. University of Chicago Press

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66: 315–331

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4): 475–492

Heckman, J.J., **Ichimura, H.**, and **Todd, P.E.** (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4): 605–654

Ho, D. and **Imai, K.** (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American Statistical Association*, 101: 888–900

Ho, K. and **Rosen, A.M.** (2017). *Partial Identification in Applied Research: Benefits and Challenges*, volume 2 of *Econometric Society Monographs*, 307359. Cambridge University Press. doi:10.1017/9781108227223.010

Hodges, J. and **Lehmann, E.** (1963). Estimates of location based on ranks. *Annals of Mathematical Statistics*, 34: 598–611

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960

Honor, B.E. and **Lleras-Muney, A.** (2006). Bounds in competing risks models and the war on cancer. *Econometrica*, 74(6): 1675–1698

Horowitz, J. and **Manski, C.F.** (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 77–84

Imbens, G.W. and **Kalyanaraman, K.** (2012). Optimal bandwidth choice for the Regression Discontinuity estimator. *Review of Economic Studies*, 1–27

Imbens, G.W. and **Lemieux, T.** (2008). Regression Discontinuity Designs: A guide to practice. *Journal of Econometrics*, 142: 615–635

Imbens, G. (2018). Understanding and misunderstanding randomized controlled trials: A commentary on deaton and cartwright. *Social Science and Medicine*, 210: 50 – 52. ISSN 0277-9536. doi:https://doi.org/10.1016/j.socscimed.2018.04.028. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue

Imbens, G.W. and **Angrist, J.D.** (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2): 467–475

Imbens, G.W. and **Manski, C.F.** (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72: 1845–1857

Imbens, G.W. and **Rubin, D.B.** (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press

Kitagawa, T. (). A test for instrument validity. *Econometrica*, 83(5): 2043–2063. doi: 10.3982/ECTA11974

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70: 1781–1803

Kreider, B. and **Pepper, J.V.** (2007). Disability and employment: reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association*, 102: 432–441

Kreider, B., **Pepper, J.V.**, **Gundersen, C.**, and **Jolliffe, D.** (2012). Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. *Journal of the American Statistical Association*, 107(499): 958–975

Kubicka, L., **Matejcek, Z.**, **Dytrych, Z.**, and **Roth, Z.** (2001). Iq and personality traits assessed in childhood as predictors of drinking and smoking behaviour in middle-aged adults: a 24-year follow-up study. *Addiction (Abingdon, England)*, 96: 1615–1628. ISSN 0965-2140. doi:10.1080/09652140120080741

LaLonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4): 604–620

Lee, D. and **Card, D.** (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142: 655–674

Lee, D. and **Lemieux, T.** (2010). Regression discontinuity desings in economics. *Journal of Economic Literature*, 48: 281–355

Lee, D.S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3): 1071–1102

Lochner, L. and **Monge-Naranjo, A.** (2012). Credit constraints in education. *Annual Review of Economics*, 4(1): 225–256. doi:10.1146/annurev-economics-080511-110920

Manski, C.F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, 80: 319–323

Manski, C.F. (1997). Monotone treatment response. *Econometrica*, 65: 1311–1334

Manski, C. (2003). *Partial Identification of Probability Distributions: Springer Series in Statistics*. Springer

Manski, C.F. (2007). *Identification for prediction and Decision*. Harvard University Press

Manski, C.F. (2018). The lure of incredible certitude. NBER working paper 24905, National Bureau of Economic Research

Manski, C.F. and **Pepper, J.V.** (2000). Monotone Instrumental Variables: with an application to the returns to schooling. *Econometrica*, 68: 997–1010

Manski, C.F. and **Pepper, J.V.** (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29(1): 123–141

Manski, C.F. and **Pepper, J.V.** (2017). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics*, 0(ja): null

Marschak, J. and **Andrews, W.H.** (1944). Random simultaneous equations and the theory of production. *Econometrica*, 12(3/4): 143–205. ISSN 00129682, 14680262

Mazzonna, F. and **Peracchi, F.** (2012). Ageing, cognitive abilities and retirement. *European Economic Review*, 56: 691–710

Mikusheva, A. (2010). Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2): 236 – 247

Milligan, K. and **Stabile, M.** (2011). Do child tax benefits affect the well-being of children? evidence from canadian child benefit expansions. *American Economic Journal: Economic Policy*, 3(3): 175–205. doi:10.1257/pol.3.3.175

Moreira, M.J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4): 1027–1048

Morgan, S.L. and **Winship, C.** (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2 edition. doi:10.1017/CBO9781107587991

Munnell, A., **Webb, A.**, and **Chen, A.** (2016). Does socioeconomics status lead people to retire too soon? Technical report, Center for Retirement Research at Boston College, Paper IB 16-14.

Nelson, C. and **Startz, R.** (1990a). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, 63(1): S125–40

Nelson, C.R. and **Startz, R.** (1990b). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, 58(4): 967–976. ISSN 00129682, 14680262

Neuman, K. (2008). Quit your job and get healthier? The effect of retirement on health. *Journal of Labor Research*, 29: 177–201

Neyman, J., **Iwaszkiewicz, K.**, and **Kolodziejczyk, S.** (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2): 107–180. ISSN 14666162

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625. ISSN 09528385

Neyman, J., **Dabrowska, D.M.**, and **Speed, T.P.** (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statist. Sci.*, 5(4): 465–472. doi:10.1214/ss/1177012031

Okumura, T. and **Usui, E.** (2014). Concave-monotone treatment response and monotone treatment selection: With an application to the returns to schooling. *Quantitative Economics*, 5(1): 175–194

Oreopoulos, P. (2006). Estimating average and local average treatement effects of education when compulsory schooling laws reall matter. *American Economic Review*, 96: 152–175

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statist. Surv.*, 3: 96–146. doi:10.1214/09-SS057

Pearl, J. (2009b). *Causality*. Cambridge University Press. doi:10.1017/CBO9780511803161

Pearl, J. (2018). Challenging the hegemony of randomized controlled trials: A commentary on deaton and cartwright. *Social Science and Medicine*, 210: 60 – 62. ISSN 0277-9536. doi:https://doi.org/10.1016/j.socscimed.2018.04.024. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue

Poterba, J.M., **Venti, S.F.**, and **Wise, D.A.** (1995). Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1): 1 – 32

Raudenbush, S.W. (2018). On randomized experimentation in education: A commentary on Deaton and Cartwright, in honor of Frederick Mosteller. *Social Science and Medicine*, 210: 63 – 66. ISSN 0277-9536. doi:https://doi.org/10.1016/j.socscimed.2018.04.030. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue

Rosenbaum, P. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, 91: 465–468

Rosenbaum, P. and **Imbens, G.W.** (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society, A*, 168: 109–126

Rosenbaum, P. (2010). *Design of observational studies*. Springer Series in Statistics. Springer-Verlag New York

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66: 688–701

Rubin, D.B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371): 591–593. ISSN 01621459

Rubin, D.B. (1990a). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3): 279 – 292. ISSN 0378-3758. doi:https://doi.org/10.1016/0378-3758(90)90077-8

Rubin, D.B. (1990b). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, 5(4): 472–480. doi: 10.1214/ss/1177012032

Schochet, P.Z., **Burghardt, J.**, and **McConnell, S.** (2008). Does job corps work? impact findings from the national job corps study. *The American Economic Review*, 98(5): 1864–1886

Senn, S. (2013). Seven myths of randomisation in clinical trials. *Statistics in Medicine*, 32(9): 1439–1450

Shah, M. and **Steinberg, B.M.** (2017). Drought of opportunities: Contemporaneous and long term impacts of rainfall shocks on human capital. *Journal of Political Economy*, 125(2): 527–561

Staiger, D. and **Stock, J.** (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65: 557–586

Stautz, K., **Pechey, R.**, **Couturier, D.L.**, **Deary, I.J.**, and **Marteau, T.M.** (2016). Do executive function and impulsivity predict adolescent health behaviour after accounting for intelligence? findings from the alspac cohort. *PloS one*, 11: e0160512. ISSN 1932-6203. doi:10.1371/journal.pone.0160512

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2(1): 167–195

Thistlethwaite, D. and **Campbell, D.** (1960). Regression-Discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51: 309–317

Wooldridge, J. (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT University Press