Università degli Studi di Milano - Bicocca

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Dottorato di Ricerca in Statistica XXIII ciclo

A model for the evaluation of graduates' first long-term job on labour market history

Discussant:
Isabella Romeo

Matricola 051724

Tutor:
Prof. Giorgio Vittadini

Co-Tutor:
Prof. Fulvia Pennoni

Academic Year 2009 – 2010

To my parents

Acknowledgements

I would first like to express my gratitude to my supervisor, Professor Giorgio Vittadini for his continuous encouragement and invaluable suggestions during this work. I would also like to extend my gratitude to Fulvia Pennoni who made many valuable suggestions and gave constructive advice that improved the quality of this study.

I would like to thank Professor Jaap Abbring for giving me the possibility to go to his department of Econometrics in Tilburg. Our discussions have been of great help for new ideas. There I met many people who gave helpful suggestions, for which I would like to thank them. Among them Matteo Picchio and Patrick Arni. Whenever I had doubts and questions, their doors were always opened.

I am also grateful to Professor Mario Mezzanzanica for continuously giving me a boost to carry on my studies with the Phd. I thank him also for providing me with interesting data to work on.

My gratitude also goes to my colleagues, particularly to Viviana, for sharing doubts, problems and difficult steps of our doctoral studies.

I would like to show my gratitude to Willy, for his constructive comments on this thesis.

My special appreciation goes to my parents, Franca and Bruno, for their tireless support throughout my undergraduate and postgraduate years of study. I am in debt to my parents for inculcating me the values of dedication and discipline to try to do well whatever I undertake. I am also thankful to my sister, Valentina, for her ever-present support.

Finally, I would like to express special thanks to my husband Massimo. He helped me concentrate on completing this dissertation and supported me mentally during the course of this work. Without his help and encouragement, this study would not have been completed.

Contents

List of Figures

List of Tables

Introduction

The goal of most scientific investigation is to uncover causal relationships. Although many standard inferential procedures may be able to conclude that an observed association between two variables is not simply due to chance, they cannot in general say whether the relationship is causal. If two variables are found to co-exist more often than an ordinary chance, then it is useful to consider the correlation between these variables. The trouble is that, unless they are properly controlled for, there could be other variables affecting this relationship that the researchers do not know about. Researchers have at their disposal a number of sophisticated statistical tools to control for these, ranging from the relatively simple one, like multiple regression, to the highly complex and involved one as the multi-level modeling. These methods allow researchers to separate the effect of one variable from others, thereby leaving them more confident in making assertions about the true nature of the relationships they found. Still, even under the best analysis circumstances, correlation is not the same as causation.

Causal inference attempts to uncover the generating structure of the data and eliminate all non-causative explanations for an observed association. An explicit introduction of the philosophy and approaches to causation was brought to the statistical literature in 1986 by Paul Holland, although references to causal approaches exist in literature up to 60 years before. The statistics literature originates in the analysis of epidemiology randomized controlled experiments by Fisher (1925) and Neyman (1990). Rubin (1973a,b, 1974, 1977, 1978), in a series of papers, formulated the now dominant approach to the analysis of causal effects in observational studies. He proposed the interpretation of causal statements as comparisons of so-called potential outcomes: pairs of outcomes defined for the same unit given different levels of exposure to the treatment. Parallel frameworks were independently developed in psychometrics (Thurstone, 1927) and econometrics (the potential outcome framework is already present in Haavelmo's (1943) work on simultaneous equation model and appear explicitly in labour market setting by Roy (1951) and Quandt (1972)).

The concept of causality developed in econometrics and in the statistical treatment effect literature is based on the notion of controlled variation that means variation in treatment holding other factors constant. There are some

authors that have considered a variety of ways in which probabilistic and causal models can be represented in graphical form. Among them see Dawid (2002) and Pearl (1995). There are other authors that use distinct notions of causality. Granger (1969) and Sims (1972) used a notion of causality based on prediction, Cartwright (2004) discussed a variety of definition from a philosopher's perspective. A useful distinction and a comparison among the commonly invoked definition of causality can be found in Holland (1986) and Lechner (2006), respectively. The recent theoretical literature has built on combined features of earlier work in both statistics and econometrics literature.

Even if the statistical and econometric literature starts from different perspective, it studies the same central problem: evaluating the effects of the exposure of a set of units to a program or treatment on some outcome of interest. A program can be made by more and different treatments or by different levels of the same treatment. The simplest case is the one where there are only treatment and non-treatment. In such a case the members of the population who take part in the program will be referred to as participants (or exposed, treated), while those who do not take part in the program are non-participants (or non-exposed, non-treated). Members of the population can be assigned or self-assigned and can be denied or self-denied the treatments according to the assignment mechanism that represents the eligibility rule and the process that decide which units receive the treatment and which do not. For example, an individual may or may not enrol in a training program, may or may not receive a voucher, may or may not be subject to a peculiar regulation. The key cognitive question is if participation to the program has any causal effect on the observed outcome of the population members, where an outcome variable is an observable characteristic or some particular measurement of the units of the population, on which the intervention may or may not apply and may have an effect or impact. The object of interest, the causal effect, is a comparison of the two outcomes for the same unit when exposed, and when not exposed, to the treatment. Most of the times this is not possible as only one of the outcomes can be observable on the same unit. This involves a problem of missing values, also known as the "fundamental problem of causal inference" (Holland, 1986). In this context a central ingredient to the definition of causal effect is the term "counterfactual". Counterfactual does not need to be contrary to certain facts, but it is just something hypothetical, imagined or nonfactual. It represents all the possible outcomes that could be verified. For example, in a cause relation between events, I can say the following counterfactual statement: "if A had occurred, then B would have occurred", even though A in fact had not occurred. Thus, in order to evaluate the effect of the treatment I need to compare distinct units receiving the different levels of the treatment, or the same unit in different moments.

In a series of papers Heckman compares the economic and statistical

approach to policy analysis (Heckman, 2005, 2008, 2010) trying to build a bridge between these approaches making the implicit economics of statistical approach explicit. In this way he extends the interpretability and range of policy questions that these methods can answer. Heckman in his last paper Heckman (2010) explains how these two approaches come out from the perceived failure of econometric structural methods. In economics evaluation of policy was based on "structural models" because the parameters of such models can answer a wide variety of policy questions. Many studies about sensitivity of estimates of these models to assumptions about functional forms and distributions on unobservables show the fragility of these estimates. Among the papers that demonstrate that standard structural estimation methods applied to non experimental data cannot duplicate the estimates obtained from a pre-program experiment, there is the one by LaLonde (1986). This and other studies produced two different responses: retreat to statistics away from the use of explicit economic models for the so called "program evaluation approach" and development of a more robust version of structural approach. In other words: statistical and econometric approaches. The focus in the econometric literature is traditionally on endogeneity, or self-selection, issues. Individuals who choose to enrol in a training program are by definition different from those who choose not to enrol. These differences, if they influence the response, may invalidate causal comparisons of outcomes by treatment status, possibly even after adjusting for observed covariates. Consequently, many of the initial theoretical studies focused on the use of traditional methods for dealing with endogeneity, such as fixed effect methods from panel data analyses and instrumental variables methods. Econometrics literature uses semiparametric and nonparametric literature to develop new estimators requiring fewer functional forms and tries to figure out heterogeneity modelling the unobservable. The complex computational method required to implement this approach makes it less transparent, replication and sensitivity analysis more difficult. Angrist and Pischke (2009) describes this approach as complex and not credible. Instead they are more favourable to the program evaluation approach, where parameters of interests are defined as summary of the output of experimental intervention, rather than through clearly explicit models. In fact the statistical literature abandons the economic choice theory, so the distinction between ex ante and ex post outcome and subjective and objective evaluations.

I apply a dynamic version of causal model in the context of the labour market, given that I have administrative panel data at my disposal. Dynamic models have been recently proposed in literature to face the fact that a treatment or a policy may be evaluated dynamically on time. Furthermore these models allow to control for unobserved heterogeneity and to estimate state dependence. Having at disposal administrative panel data on both Lombardy labour market and records of the graduates of three biggest Uni-

versity of Milan, it is of great interest to use such models to study the impact of the first "stable" job coherent with the university education on the future job coherence. To the best of my knowledge there are no papers that focus on job coherence. Moreover, most articles assume a permanent job as stable and focus on the time to get it, given that subjects have a temporary job or are unemployed (see for example Gagliarducci (2005); Bonnal *et al.* (1997); Gritz (1993)). The global economic system has changed and with it business and workers needs. People lose the certainty of a life-long lasting workplace and a career that develops within a company with a known location and well defined growing path. Given the increased instability of the market, a new concept of work has arisen: the *work path*. In such a context the permanent contract cannot be considered a stable one anymore (Bauer and Bender, 2004). From here rises the interest of defining the concept of stability and of understanding what happens after getting a stable job. Furthermore, having information also on university career allows to consider also the coherence job in the definition of a good job.

The first part of the present work attempts to sum up reviews and discussions about causal inference from the statistical and econometric literatures, describing and comparing the two main approaches. In the first chapter an introduction on the concept of causality and the main methods to make causal inference without explicit models are presented. This represents a review of the statistical literature following the principal surveys of Imbens and Wooldridge (2009), Pearl (2010), Caliendo (2006), Angrist and Pischke (2009), Rosenbaum (1995) and Pearl (2000). The second chapter focuses on the econometric approach, or rather on methods to make causal inference with explicit models with a comparison to the program evaluation approach. The main features of this approach are described following the overview of the important theoretical work by Heckman and his co-authors in the econometric literature. The third chapter proposes a survey of the main dynamic models used in the statistical and econometric literature following the survey of Abbring and Heckman (2008) and Robins (Robins, 1989, 1997; Robins *et al.*, 2000; Gill and Robins, 2001). In the second part of my work I attempted to use such models to study the labour market in Lombardy, focusing on stability and job coherence with the university studies. Chapter four introduces the labour market in Italy and describes the data at disposal through descriptive analysis. Finally, in the last chapter I presented the model used and the results about the impact of the first "stable" job coherent with the university studies on the future job coherence.

Chapter 1

Causation and potential outcome approach

Many empirical questions in economics and other social sciences depend on causal effects of programs or policies. In the last two decades, much research has been done on the econometric and statistical analysis of the effects of such programs or treatments. A first concept to clarify is the definition of causation and its distinction from association. It is important to avoid confusion between these two different concepts not to run into a fallacy. The potential outcome framework is useful to give a definition of causality, given that it is the base of causal inference. This framework is known in literature as the Newman-Rubin approach or statistical approach. This chapter synthesizes the main features of the potential outcome model and the main parameters of interest considered in this literature. It illustrates the principal techniques of causal inference based on the assumptions of unconfoundedness and of unobservables without the explicit use of statistical models. Among the model under the assumptions of unconfoundedness the matching technique is analyzed, while among the model of unobservable the instrumental variables, regression discontinuity design and difference-and-difference methods are presented.

1.1 From association to causation

In the last decades lots of attention has turned on causality. It is a common mistake to assume cause and effect for two variables simply because they occur together. In such a way a statistical correlation is given a causal interpretation. A known example in literature is the following. During the 1990s both religious attendance and illegal drug use were on the rise: it would be erroneous to conclude that therefore, religious attendance causes illegal drug use. It is also possible that drug use leads to an increase in religious attendance, or that both drug use and religious attendance are

increased by a third variable, such as an increase in societal unrest. It is also possible that both variables are independent one from the other, and it is a mere coincidence that they are both increasing at the same time. The problem with assuming cause and effect from mere correlation is not that a causal relationship is impossible; it is just that there are other variables that must be considered and not ruled out a-priori.

To understand the cause-effect relation it is necessary to understand the concept of association well, to avoid the confusion between them. An association is said to exist between two variables when a change in one variable coincides with a change in another. This is also called *covariation* or *correlation*. An association may be positive or negative and may be proportionate or disproportionate. There are various factors which may explain why an association can be spurious:

- *Chance* may have affected results because of random variation in the population. It could be that, by chance, the sample chosen it is representative of a peculiar subpopulation. In that case, underestimation or overestimation of the effect may occur.

- In some aspect of the design, or during the study, some systematic errors or bias could be introduced into the results: *Selection Bias* and *Observable Bias*. The former occurs if the study populations being compared are not strictly comparable and the latter if non-comparable information is obtained from each study group.

- A third, and most important, possibility is *confounding*. A confounding is a variable that distorts the relationship between other two variables, because it is correlated with both of them. It can obscure the relationship of interest of spuriously create one.

Furthermore association implies a symmetric relation between variables: if a variable X is correlated with Y then also Y is correlated with X. A causal relation has a defined direction: if X causes Y, it cannot be that also Y causes X. To understand confounding, let us assume that a variable X is correlated with a variable Y. X may not be a cause of Y but rather X and Y may have a common cause Z, which accounts for their correlation. Or X may be an effect of Y. Figure 1.1 shows these possibilities for the causal basis for a correlation between the two variables.

It is very easy to make mistakes proving causality. This could have disastrous consequences if the errors form the basis of public policy. This is nothing new. Freedman (1999) describes one of the earliest attempts to use regression models in the social sciences. Yule (1899) investigated the causes of pauperism in England. Depending on local custom, paupers were supported inside local poor-houses or outside. Yule used a regression model to analyze his data and found that the change in pauperism was positively

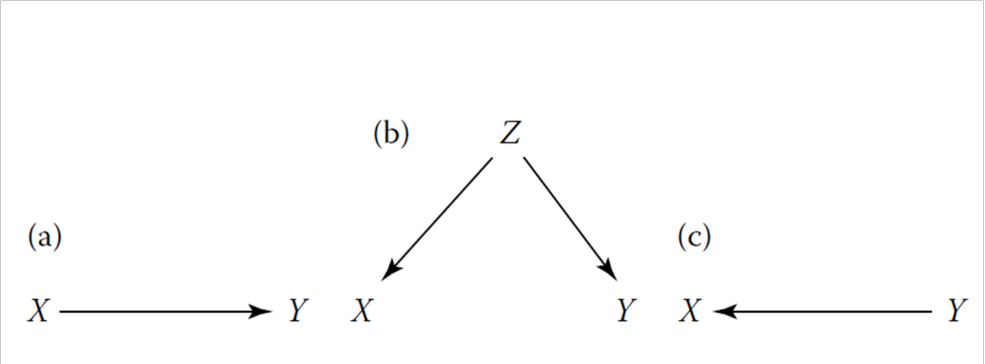

Figure 1.1: Three forms of causation that produce a correlation between X and Y

related to the change in the proportion treated outside poor-houses. He then reported that welfare provided outside poor-houses created paupers. A contemporary of Yule's suggested that what Yule was seeing was instead an example of confounding: those areas with more efficient administrations were better at both building poor-houses and reducing poverty. That is, if efficiency could be accounted for, there would be no association between pauperism and the way aid was provided. Freedman notes that after spending much of the paper assigning parts of the change in pauperism to various causes, Yule left himself out with his footnote : "Strictly speaking, for "due to" read "associated with" ". Cause and effect is established through an experiment in which two groups participate at the same experience except for a single factor. Any difference in outcome is then attributed to that factor.

The purpose of an observational study is to elucidate cause-and-effect relationships. The most familiar difficulty is that, since treatments were not randomly assigned to experimental units, treated and control groups may not be directly comparable. Even after adjustments have been made for observed covariates, estimates of treatment effects can still be biased by imbalances in unobserved covariates treatments.

1.1.1 An informal definition of causal effect

A simple example is now proposed to intuitively understand the definition of causal effect. Let us consider a dichotomous treatment variable A that assumes value 1 if the individual is treated and 0 otherwise. Let $Y(1)$ be the outcome variable that would have been observed under the treatment $(A = 1)$ and $Y(0)$ be the outcome variable that would have been observed under no treatment $(A = 0)$. Intuitively it is possible to say that the treatment as a causal effect is $Y(1) \neq Y(0)$.

From the observational data it is not possible to observe both $Y(0)$ and

$Y(1)$ for the same subject. For each subject is observed only the treatment level A and the observed outcome Y. The available data can be used to define the conditional probability $P(Y = 1|A = 1)$ as the proportion of subjects that have the outcome $Y = 1$ in the population of interest that receive treatment A. When the proportion of subjects who have the outcome $Y = 1$ in the treated $P(Y = 1|A = 1)$ equals the number of subjects who have the outcome $Y = 1$ in the untreated $P(Y = 1|A = 0)$, it is possible to say that treatment A and outcome Y are *independent*. Independence is a symmetric relation and implies that A is not associated with Y, or that A does not predict Y. Some equivalent definitions of *independence* are:

- risk difference $P(Y = 1|A = 1) - P(Y = 1|A = 0) = 0$;
- risk ratio $\dfrac{P(Y = 1|A = 1)}{P(Y = 1|A = 0)} = 1$;
- odds ratio $\dfrac{P(Y = 1|A = 1)/P(Y = 0|A = 1)}{P(Y = 1|A = 0)/P(Y = 0|A = 0)} = 1$.

Treatment A and outcome Y are *dependent* if the above equations do not hold. In such a case these measures quantify the strength of the association between the two variables. It is possible to rewrite the measures of association in the population as $E(Y|A = 1) \neq E(Y|A = 0)$ given that the risk equals the average in the population. Furthermore rewriting the association in this way permits to apply the definitions to non dichotomous outcomes.

The key difference between association and causation is that while a causal effect is defined as a comparison of the same subjects under different actions, instead an association is defined as a comparison of different subjects under different conditions. The risk $P(Y = 1|A = 1)$ is a conditional probability in the subset of the population that have received the treatment. In contrast the risk defined as $P(Y(1))$ is an unconditional (or marginal) probability. It represents the risk of $Y = 1$ in the entire population. Therefore, association is defined by a different risk in two disjoint populations determined by the subject's actual treatment value $(A = 1, A = 0)$, whereas causation is defined by a different risk in the entire population under two different treatment values.

In general "association is not causation" and usually because those who were treated and those who were untreated in the population must be different in some sense. Causal inference requires data like the hypothetical data, so the question is under which conditions real world data can be used for causal inference. In the following paragraphs some methods are presented to find causal effect based on the potential outcome approach without using explicit models (the so called "statistical approach"). In the following chapter there is instead a review of methods with explicit models (the so called "econometric approach").

1.2 Potential outcome approach

The potential outcome approach (POA) was labeled as the Rubin Causal Model (RCM) by Holland in 1986 (Holland, 1986), but its origin goes back to Cox (1972) and Neyman (1990). This approach is by now standard in statistical literature. See Imbens and Wooldridge (2009) for a more detailed review.

1.2.1 Definitions and main assumptions

All the examples used in this paper will deal with an economic context, given that the application of Chapter 5 is on labour market. Let us start with an example such as the analysis of a job training program. There is information about N individuals. Some of them were enrolled and others were not, either because they were ineligible or chose not to enroll. The indicator used for the assignment is $D(\omega)$ to indicate whether individual ω enrolled in the training program ($D(\omega) = 1$) or not ($D(\omega) = 0$). Let N_0 and N_1 denote the number of control and treated units, respectively. For each unit it is also observed a K-dimensional column vector of covariates or pre-treatment variable, $X(\omega)$, with X denoting the $N \times K$ matrix with ω-th row equal to $X(\omega)'$.

For individual ω, for $\omega = 1, \ldots, N$, is postulated the existence of two potential outcomes, denoted by $Y(\omega, 0)$ and $Y(\omega, 1)$. The first, $Y(\omega, 0)$, denotes the outcome that would be realized by individual ω if he or she did not participate in the program ($D(\omega) = 0$). Similarly, the second, $Y(\omega, 1)$, denotes the outcome that would be realized by individual ω if he or she participated in the program ($D(\omega) = 1$). Individual ω can participate or not in the program, but not both, and thus only one of these two potential outcomes can be realized. Prior to the assignment being determined, both are potentially observable, hence the label potential outcomes. If individual ω participates in the program $Y(\omega, 1)$ will be realized and $Y(\omega, 0)$ will ex post be a counter-factual outcome. The realized outcome is denoted by $Y(\omega)$ and with Y the N-vector with ω-th element equal to $Y(\omega)$. The preceding discussion implies that

$$Y(\omega) = Y(\omega)D(\omega) = Y(\omega, 0)(1 - D(\omega)) + Y(\omega, 1)D(\omega) =$$

$$= \begin{cases} Y(\omega, 0) & \text{if } D(\omega) = 0 \\ Y(\omega, 1) & \text{if } D(\omega) = 1. \end{cases}$$

The distinction between the pair of potential outcome $(Y(\omega, 0), Y(\omega, 1))$ and the realized outcome $Y(\omega)$ is the hallmark of modern statistical and econometric analyses of treatment effect.

The most common definition of the casual effect at the unit level is the difference $Y(\omega, 1) - Y(\omega, 0)$, but it is possible to consider also the ratio

$Y(\omega,1)/Y(\omega,0)$ or other functions.

Such definition does not require to take a stand on whether the effect is constant or varies across the population. Furthermore, defining individual specific treatment effect using potential outcomes does not require to assume endogeneity or exogeneity of the assignment mechanism. By contrast, the casual effects are more difficult to define in terms of the realized outcomes. The regression function of realized outcome $Y(\omega) = \alpha+\tau+D(\omega)+\epsilon(\omega)$ is then interpreted as a structural equation with τ casual effect. Leaving unclear whether the casual effect is constant or not, and what the property of the unobserved component $\epsilon(\omega)$ is. Considering also some covariates $X(\omega)$ and supposing that the treatment effect is constant $\tau = Y(\omega,1)-Y(\omega,0)$ then the regression function of realized outcome becomes $Y(\omega,0) = \alpha+\beta' X(\omega)+\epsilon(\omega)$. The residual can be written as $\epsilon(\omega) = Y(\omega,0)-\mathbb{E}[Y(\omega,0)|X(\omega)]$. It captures the unobservable affecting the response in the absence of the treatment. The observed outcome is instead defined as:

$$Y(\omega) = (1 - D(\omega)) \cdot Y(\omega,0) + D(\omega) \cdot Y(\omega,1) = \alpha + \tau \cdot D(\omega) + \beta' X(\omega) + \epsilon(\omega).$$

The potential outcomes approach allows the researchers to first define the casual effect of interest without considering probabilistic properties of the outcomes or assignments.

Intuitively to individuate causal effects there should not be any "contagion" between members of the target-population, either with regard to the (self) allocation to the treatment state or to the outcome(s). Moreover, in order to make comparisons, it would be necessary to observe both treated and non-treated members and they should be sensibly comparable, that is, they should not differ with respect to characteristics that affect both the self(allocation) to the treatment state and the outcome, except for the fact that some members are treated and some others are not.

In most of the literature it is assumed that the treatment received by one unit does not affect outcomes for another unit. Only the level of the treatment applied to the specific individual is assumed to potentially affect outcome for that particular individual. In statistic literature this assumption is referred to as the Stable-Unit-Value-Assumption (SUTVA, Rubin (1978)). This means treatment and no treatment should be well defined, and stable across units. Furthermore it embodies the crucial assumption of no interactions between members of the target population, in the sense that the treatment state and the outcome experienced by member is not affected by the treatment state not by the outcome of any other member of the population. Formally:

$$(Y(\omega_1,1), Y(\omega_1,0), D(\omega_1,)) \perp\!\!\!\perp (Y(\omega_2,1), Y(\omega_2,0), D(\omega_2,)) \text{ for all } \omega_1 \neq \omega_2.$$

This lack of interaction assumption is very plausible in many biomedical applications. However there are also many cases in which such assumption

is not plausible. In economic applications, interaction between individuals is a serious concern. One general solution to this problem is to redefine the unit of interest. If the interaction between individuals is at an intermediate level, say a local labor market, one can analyze the data using the local labor market as the unit and changing the no-interaction assumption to require the absence of interactions among local labor. Such aggregation is likely to make the no-interaction assumption more plausible. An alternative solution is to directly model the interactions. This involves specifying which individuals interact with each other, and possible relative magnitudes of these interactions (in some cases it may possible to assume that interactions are limited to individuals well defined or interactions occur in broader groups but decline in importance depending on some distance metric).

In many early studies it was assumed that the effect of a treatment was constant, implying that the effect of various policies could be captured by a single parameter. The essentially unlimited heterogeneity in the effects of the treatment allowed for in the current literature implies that it is generally not possible to capture the effects of all policies of interest in terms of a few summary statistics. Most of the estimands are average treatment effects, although some correspond to other features of the joined distribution of potential outcomes. Most of the empirical literature has focused on estimation. Much less attention has been devoted to testing hypotheses regarding the properties or presence of treatment effects.

1.2.2 Main advantages

The definition of causal effect according to the potential outcome framework of Rubin has several remarkable advantages (Imbens and Wooldridge, 2009; Heckman and Vytlacil, 2007b):

- It allows one to define causal effects without making functional form and/or distributional assumptions. Of particular importance in Rubin's approach is the relationship between treatment assignment and the potential outcomes. The simplest case is when assignment to treatment is randomized, and thus independent of covariates as well as the potential outcomes. In such cases it is straightforward to obtain attractive estimators for the average effect.

- The unit level causal effect is free to vary across units, it allows general heterogeneity in the effect of the treatment: some population members benefit more from the intervention, some others benefit less, some others might even be damaged by the intervention.

- It links the analysis (or identification) of causal effects to explicit manipulations. Considering the two potential outcomes forces the analyst to think about scenarios under which each outcome will be observed,

that is considering the kinds of (randomized or natural) experiments that could reveal the causal effects.

- Specifically, it does not rely on the way in which the treatment state is determined (the treatment state might result from individual choices, the external decision of an authority or the toss of a coin). Thus, it separates the potential outcomes from the modeling of the assignment mechanism. Modeling the realized outcomes is complicated by the fact that it "collapses" the potential outcomes and the assignment mechanism.

- It allows to formulate probabilistic assumptions in terms of potentially observable variables, rather than in terms of unobserved components. Many of the critical assumptions will be formulated as (conditional) independence assumptions involving the potential outcomes. Models specified in terms of realized outcomes often formulate the critical assumptions in terms of errors in regression functions. In such way a they implicitly bundle a number of assumptions, as functional form and exogeneity, difficult to assess.

- It clarifies where the uncertainty in the estimators comes from. Even if the entire population is observed casual effects will be uncertain because for each unit at most one of the two potential outcomes is observed.

1.2.3 Average treatment effects and other features of the distribution of causal effects

The definition of causal effect is based on the difference of potential outcomes for each member of the target-population. The unit level difference by its very nature cannot be observed, as it involves a counter-factual outcome. As a consequence, also the distribution of the causal effect, as well as any summary measure of it (such as the population level causal effect), cannot be observed. A possible solution is to focus on specific features of the distribution of the causal effect in the target-population (or suitable subsets of it), chiefly means, and on their identification and estimation.

The literature has largely focused on average effects of the treatment. Since the expected value is a linear operator, to evaluate *average treatment effect (ATE)* it is enough to observe $E[Y(1)]$ and $E[Y(0)]$:

$$E[Y(\omega,1) - Y(\omega,0)] = E[Y(\omega,1)] - E[Y(\omega,0)].$$

Another popular estimand is the *Average Treatment effect on the Treated (ATT)*, the average over the subpopulation of treated units:

$$\tau_{att} = \mathbb{E}[Y(\omega,1) - Y(\omega,0)|D(\omega) = 1].$$

In many observational studies τ_{att} is a more interesting estimand than the overall average effect.

However, ATE and ATT are not observed, but the factual outcome can be observed for the treated after treatment $E[Y(1)|D = 1]$ and for the untreated after not receiving treatment $E[Y(0)|D = 0]$. Subtracting and adding, to the observed difference in means between treated and non-treated, the counter-factual outcome for the treated units $E[Y(0)|D = 1]$, you get the key identity:

$$E[Y(1)|D = 1] - E[Y(0)|D = 0] =$$

$$= E[Y(1)|D = 1] - E[Y(0)|D = 1] + E[Y(0)|D = 1] - E[Y(0)|D = 0].$$

The first term of the equation $E[Y(1)|D = 1] - E[Y(0)|D = 1]$ represents the Average Treatment effect on the Treated (ATT). It consists of the difference between the factual outcome for the treated after treatment, and their counter-factual outcome (the outcome we would have observed for the treated, had they not been exposed to treatment).

The second term $E[Y(0)|D = 1] - E[Y(0)|D = 0]$ represents the *selection bias*. This is the outcome difference that would be observed between participants and non-participants if the programme was not implemented, and it depends on pre-existing differences between the two groups. That is, this term captures outcome differences between participants and non-participants that cannot be attributed to the programme.

The term $E[Y(1)|D = 1] - E[Y(0)|D = 0]$ is the observed mean difference between treated and non-treated, that is, between their factual outcomes, respectively. It cannot be given a causal interpretation in general, as it is the sum of mean differences that are solely induced by the programme (the ATT) and mean differences that would have occurred even in the absence of the programme (the selection bias). Whether or not this term can be given a causal interpretation crucially depends on having no selection bias. Sadly enough, the selection bias comprises a term which is observable, $E[Y(0)|D = 0]$, and a term which is not observable, $E[Y(0)|D = 1]$, the latter involving a counter-factual outcome for participants. Thus, it follows that without additional information/assumptions any causal conclusion drawn from $E[Y(1)|D = 1] - E[Y(0)|D = 0]$ is precluded because the selection bias can not be measured.

In such case the probability to participate in the intervention depends on a set of characteristics X of the population members or characteristics that refer to the social context in which the intervention takes place. They may be interpreted as the set of characteristics summarizing the selection process, namely determining how the $D = 0$ and the $D = 1$ groups are formed. This imply that the composition of the treatment and the no-treatment groups with respect to the X in general depends on D because

the selection process depends on X. It follows that if X is correlated to $Y(0)$, then $E[Y(0)|D = 1] \neq E[Y(0)|D = 0]$ because the two groups are not equivalent with respect to X. The magnitude of the selection bias depends on the selection process (or assignment mechanism), namely the set of rules (known or unknown to the analyst) according to which some members of the population are exposed to the intervention while some others are not.

If the effect of the treatment is constant $Y(\omega, 1) - Y(\omega, 0) = \tau$, τ_{cate}, τ_{ate} and τ_{att} are obviously identical. However, if there is heterogeneity in the effect of the treatment, the estimands may all be different. If there is heterogeneity, to estimate the sample average treatment effect τ_{cate} is more precise than the population average treatment effect τ_{ate}.

There are some cases in which it is difficult and uninteresting to consider the effect for the comparison group. For example, in the setting of a voluntary program, those not enrolled will never be required to participate in the program. In practice, there is typically little motivation for the focus on the overall average effect or the average effect for treated. The overall average effect would be the parameter of interest if the policy under consideration is a mandatory exposure to the treatment versus complete elimination. Similarly the average effect for the treated would be informative about the effect of entirely eliminating the current program. In this case this suggests to focus on the average casual effect conditional on the covariates in a particular sample:

$$\tau_{cate} = \frac{1}{N} \sum_{\omega=1}^{N} \mathbb{E}[Y(\omega, 1) - Y(\omega, 0)|X(\omega)].$$

Other features of the distribution of causal effects

The linear operator property does not work with other important features of the distribution of the causal effect. For example, it is not held for the median (or any other percentile) of the causal effect

$$MED[Y(1) - Y(0)] \neq MED[Y(1)] - MED[Y(0)],$$

and for the variance:

$$var[Y(1) - Y(0)] = var[Y(1)] + var[Y(0)] - 2cov[Y(1), Y(0)].$$

This parameter is particularly useful, in that it describes how treatment gains are distributed across population members. However, it can not be measured, as it requires the joint observability of the potential outcomes, which is precluded by definition.

An alternative class of estimands concerns quantile treatment effects:

$$\tau_q = F^{-1}_{Y(1)}(q) - F^{-1}_{Y(0)}(q)$$

as the q-th quantile treatment effect. These quantile effects are defined as differences between quantiles of the two marginal potential outcome distributions, rather than as quantiles of the difference:

$$\tilde{\tau}_q = F^{-1}_{Y(1)-Y(0)}(q).$$

The main reason to pay more attention on the first indicator is that $\tilde{\tau}_q$ is in general not identified without assumptions on the rank correlation between the potential outcomes.

Testing

Most of the testing in applied work has focused on the null hypothesis that the average effect of interest is zero. Because many of the commonly used estimators for average treatment effects are asymptotically normally distributed with zero asymptotic bias, it follows that standard confidence intervals can be used for testing such hypotheses. There are other interesting hypotheses to consider, for example if there is any effect of the program, that is whether distribution of $Y(\omega, 1)$ differs from that of $Y(\omega, 0)$. That is equivalent to the hypotheses that all moments are identical in the two groups.

In many cases, however, there are other null hypotheses of interest. Crump *et al.* (2008) develop a test for the null hypotheses of zero average effect conditional on the covariates:

$$H_0 : \tau(x) = 0 \quad \forall x \quad \text{against} \quad H_0 : \tau(x) \neq 0 \quad \text{for some x}$$

and a test of a constant average effect conditional on the covariates:

$$H_0 : \tau(x) = \tau_{ate} \quad \forall x \quad \text{against} \quad H_0 : \tau(x) \neq \tau_{ate} \quad \text{for some x.}$$

One may also be interested in testing the null hypothesis that the conditional distribution of $Y(\omega, 0)$ given $X(\omega) = x$ is the same as the conditional distribution of $Y(\omega, 1)$ given $X(\omega) = x$. Under the hypothesis of unconfoundedness, this is equivalent to testing the null hypothesis that:

$$H_0 : Y(\omega) \perp\!\!\!\perp D(\omega) | X(\omega)$$

against the alternative hypothesis that $Y(\omega)$ is not dependent on $D(\omega)$ given $X(\omega)$.

A second set of questions concerns treatment effect heterogeneity. It may be interesting to establish whether there are any subpopulations with an average effect positive or different from zero, or whether there are any subpopulations with an average effect exceeding some threshold. It may be also interesting to test whether there is any evidence of heterogeneity in the treatment effect by observable characteristics.

Abadie (2002) studies tests about the equivalence of outcome distributions. Such tests are studied in the settings with randomized experiments as well as settings with instrumental variables using Kolmogorov-Smirnov type testing procedures.

1.2.4 Assignment mechanism

It is possible to distinguish three main classes of assignment mechanism: randomized experiments, assignment mechanism with conditional independence and assignment mechanism with some dependence on potential outcome.

In *randomized experiments*, the probability of assignment to treatment does not vary with potential outcomes, and is a known function of the covariates. In this way the treatment state D can be assimilated to the toss of a coin, it does not systematically depend on the potential outcomes or any other characteristics. This implies that a randomized experiment is an assignment mechanism that implies zero selection bias. The observable difference $E[Y(1)|D = 1] - E[Y(0)|D = 0]$ gets equal to $E[Y(1)] - E[Y(0)]$. A randomized experiment corresponds to the assumptions of ceteris paribus and full common support. The randomized experiment is as a sort of gold standard with reference to the properties of the other identification strategies which will be assessed. All these strategies aim at reproducing as closely as possible the fundamental feature of an experimental design: having two groups equivalent in all relevant respects but different with reference to the probability of being exposed to the intervention.

The second class of assignment mechanism maintains the restriction that the assignment probabilities do not depend on the potential outcomes, or:

$$D(\omega) \perp\!\!\!\perp (Y(\omega,0), Y(\omega,1),)\,|X(\omega),$$

where $A \perp\!\!\!\perp B|C$ denotes *conditional independence* of A and B given C. However, in contrast to randomized experiments, the assignment probabilities are no longer assumed to be a known function of the covariates. Rubin (1990) refers to this assignment mechanism as *unconfounded assignment*, but in literature there are various other labels as *observables*, *exogeneity* and conditional independence. The selection bias arises in those instances in which participation depends on characteristics X which are known on a priori grounds to affect the outcome variable $Y(0)$ and are unequally distributed between participants and non-participants. Whether or not this characteristics are known and observable to the analyst is problem specific, and crucially rests upon knowledge of the selection process. The more it is known about factors determining selection process, the better it is possible to control for it by conditioning on these factors, and thus draw credible conclusions on the treatment effects. Drawing correct causal inference is primarily about ingredients knowledge of the relevant characteristics of the reference population members and the socio-economic context in which the

programme takes place, that is, knowledge of all the X that determine the selection process, and availability of adequate data on them. Statistical methods are a tool for checking the relevant assumptions, and a way to make the most out of such ingredients to come out with robust inference about program effects. A sufficient condition for the selection bias to vanish is to have the two treatment arms equally balanced with respect to all those variables that are relevant for the outcome. This is the *ceteris paribus* clause. Thus, any selection process that guarantees this condition will in general allow to draw causal inference from the observable difference. This is precisely the condition which combines CIA and common support.

The third class of assignment mechanism contains all remaining *assignment mechanism with some dependence on potential outcome*. Many of these create substantive problems for the analysis, for which there is no general solution. In such cases a possibility consists to focus on estimands that can be identified under weaker conditions than required for the average treatment. In addition there are some methods that relax the unconfoundedness assumption but do not replace it with additional assumptions. For example, it is possible to relax the unconfoundedness assumptions in a limited way and investigate the sensitivity of the estimates to such violations or to drop the unconfoundedness assumption entirely and establish bounds on estimands of interest.

1.3 Estimation and inference under unconfoundedness

As explained above, the mechanism of assignment of the program/treatment assumes great importance. The most common assumption is that the assignment probabilities do not depend on the potential outcome, that is the unconfoundedness assumptions. In the following paragraphs the main assumptions to identify the effect under unconfoundedness are described, in addition to the SUTVA condition presented in Section 1.2.1. Furthermore the way to asses this assumption and the matching technique are presented.

1.3.1 Main assumptions

Methods for estimation of average treatment under unconfoundedness are the most widely used in this literature. Unconfoundedness implies that we have sufficiently rich set of predictors for the treatment indicator, contained in the vector of covariates $X(\omega)$, such that adjusting for differences in these covariates leads to valid estimates of casual effects. The main assumption is *unconfoundedness* (also called Conditional Independence Assumption, CIA), introduced by Rosenbaum and Rubin (1983):

$$D(\omega) \perp\!\!\!\perp (Y(\omega,0), Y(\omega,1)) \,|X(\omega).$$

This assumption is controversial, as it assumes that beyond the observed covariates $X(\omega)$ there are no unobserved characteristics of the individual associated both with the potential outcomes and the treatment. The assignment mechanism depends only on observables (motivation because this assumption is also called selection on observable). Nevertheless, this kind of assumptions is used routinely in multiple regression analysis. There is a milder version relevant for causal inference on the treated. The assignment of treatment state over the support of treated $X_1(\omega)$ is independent of the unobserved potential outcomes:

$$D(\omega) \perp\!\!\!\perp Y(\omega, 0)|X(\omega)$$

for every x such that $P(D(\omega) = 1|X(\omega)) > 0$. The second assumptions used to identify treatment effects is that for all possible values of the covariates, there are both treated and control units. It is called *(Full) common support* in econometric parlance (Heckman and co-authors) and replication or *overlap* in statistical parlance (Rubin co-authors). Replication implies variability of the realized treatment arms, across units or over time. For example, if the process that decides which units receive treatment and which receive no treatment (or the assignment mechanism) depends on some observed variables X, then the assumption formally becomes

$$0 < Pr(D(\omega) = 1|X(\omega)) < 1,$$

for all x. It implies that the support of the conditional distribution of $X(\omega)$, given $D(\omega) = 0$, overlaps completely with that of the conditional distribution of $X(\omega)$ given $D(\omega) = 1$. That is, it requires a full common support for treated and non-treated units. In Figure 1.2, Case(a) is showed as the support for treated and non-treated is the same (X_1 and X_1, respectively). If the important case when the analyst aims at investigating the mean causal effect on the treated, the assumption is milder, and requires full common support of X_1 ($D(\omega) = 1$) (as shown in Figure 1.2, Case(b)), that means:

$$Pr(D(\omega) = 1|X(\omega)) < 1.$$

If the common support assumption holds partially, that is just for a subset of values of X_1, causal inference will be restricted to that subset(Figure 1.2, Case(c)). Otherwise stated, the analyst is forced "to compare only comparable people": those who share the same support. For example, in Figure 1.3 the only comparable people are those for whom there is a control group, that means people who have a X's value from 0.2 to 0.7. This assumption entails also the notions of manipulability and discontinuity. A programme can be assigned to some members of the target-population and denied to other members (discontinuity across units) or it can be introduced and canceled, or modified (discontinuity over time). See Figure 1.2, Case(d).

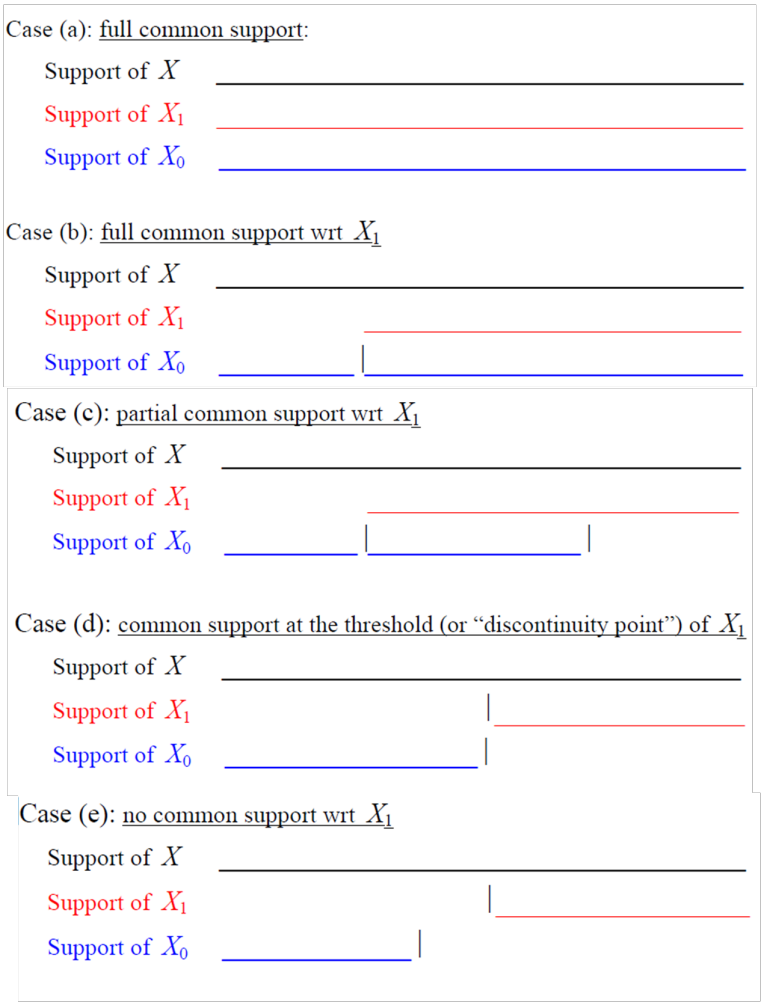

Figure 1.2: Some graphical examples of the common support issue

Finally Figure 1.2, Case(e), shows the case in which there are no common support. There is no overlap between the support of treated (X_1) and untreated (X_0). It is important to compare comparable people. Many non-experimental evaluations identify the parameter of interest by comparing observationally different persons using extrapolations based on inappropriate functional forms imposed to make incomparable people comparable. A major advantage of non-parametric methods (the potential outcome framework) for solving the problem of selection is that, rigorously applied, they force the analyst to compare only comparable people. It implies that the support of the conditional distribution of $X(\omega)$ given $D(\omega) = 0$ overlaps completely with that of the conditional distribution of $X(\omega)$ given $D(\omega) = 1$. With a random sample $(D(\omega), X(\omega))_{\omega=1}^{N}$ we can estimate the propensity score $e(x) = Pr(D(\omega) = 1|X(\omega))$ and this can provide some guidance for

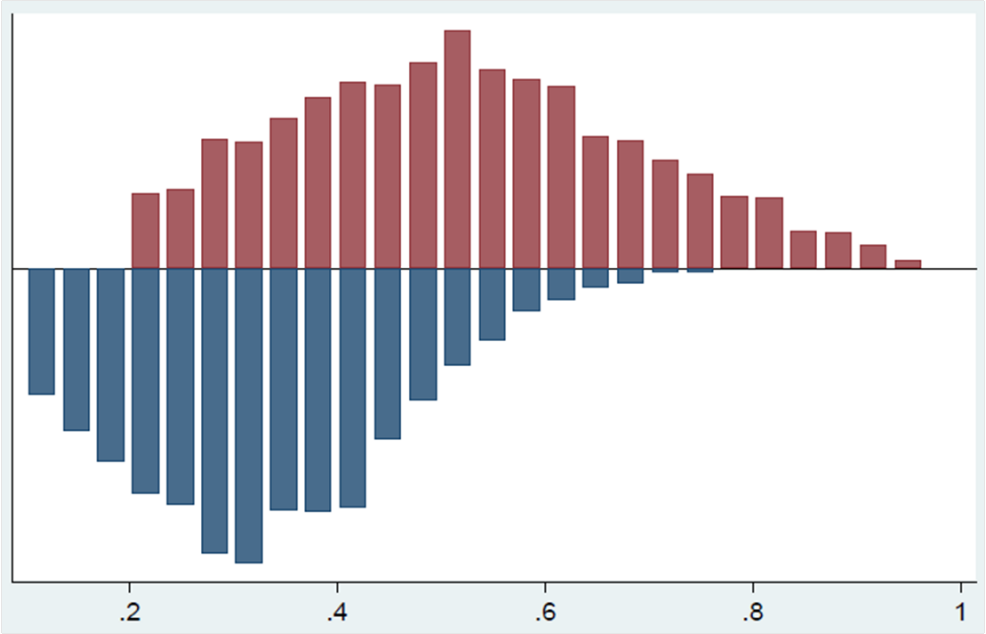

Figure 1.3: A non-parametric method for matching treated (red) and non-treated (blue)

determining whether the overlap assumptions hold. Common parametric models, such as probit and logit, ensure that all estimated probabilities are strictly between zero and one, and so examining the fitted probabilities from such models can be misleading. The combination of these two assumptions, unconfoundedness and overlap, was referred to by Rosenbaum and Rubin (1983) as *strong ignorability*. There are various ways to establish identification of various average treatment effects under strong ignorability. The easiest one is to note that $\tau(x) \equiv \mathbb{E}[Y(\omega,1) - Y(\omega,0)|X(\omega)]$ is identified for x in the support of the covariates:

$$\begin{aligned}
\tau(x) &= \mathbb{E}[Y(\omega,1) - Y(\omega,0)|X(\omega)] = \\
&= \mathbb{E}[Y(\omega,1)|X(\omega) = x] - \mathbb{E}[Y(\omega,0)|X(\omega)] = \\
&= \mathbb{E}[Y(\omega,1)|D(\omega) = 1, X(\omega) = x] - \mathbb{E}[Y(\omega,0)|D(\omega) = 0, X(\omega)] = \\
&= \mathbb{E}[Y(\omega)|D(\omega) = 1, X(\omega) = x] - \mathbb{E}[Y(\omega)|D(\omega) = 0, X(\omega)].
\end{aligned}$$

The third equality follows by the unconfoundedness $\mathbb{E}[Y(\omega)(w)|D(\omega) = d, X(\omega)]$ that does not depend on d. By the overlap assumption, it is possible to estimate both terms in the last line and identify $\tau(x)$. Identifying $\tau(x)$ for all x it is possible to identify the expected value across the population distribution of the covariates as, for example, $\tau_{ate} = \mathbb{E}[\tau(X(\omega))]$.

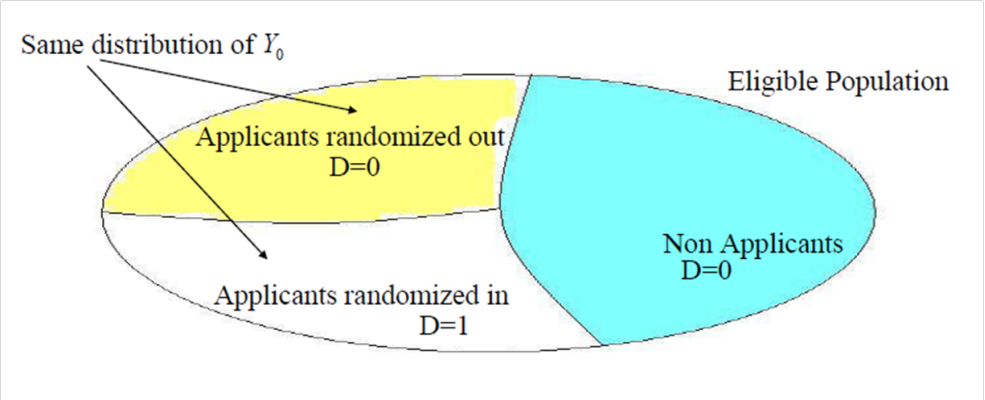

Figure 1.4: Hypothetical experiment with two control groups

1.3.2 Assessing the unconfoundedness assumption

The unconfoundedness assumption is not testable, because the data are uninformative about the distribution of $Y(\omega, 0)$ for those who received the active treatment and of $Y(\omega, 1)$ for those receiving the control. Nevertheless, there are often indirect ways of assessing this assumption. The most important of these were developed in Rosenbaum (1987) and Heckman and Hotz (1989). Both methods rely on testing the null hypothesis that an average casual effect is zero, where the particular average casual effect is known to equal zero. If the testing procedure rejects the null hypotheses, this is interpreted as weakening the support for the unconfoundedness assumption. These tests can be divided into two groups.

The first set of tests focuses on estimating the casual effect of a treatment that is known not to have an effect. It relies on the presence of two or more control groups. Supposing to have two potential control groups, for example eligible non participants and ineligibles (Figure 1.4), it is possible to estimate a *pseudo* average treatment effect by analyzing the data from these two control groups as if one of them is the treatment group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero effect implies that at least one of the control groups is invalid. Not-rejecting the test does not imply the unconfoundedness assumption is valid, but not-rejection in the case where the two treatment control groups could potentially have different biases makes it more plausible that the unconfoundedness assumption holds. Alternatively one may use geographically comparison group. Let $G(\omega)$ be an indicator variable denoting the membership of the group, taking on three values, $G(\omega) \in \{-1, 0, 1\}$. For units with $G(\omega) = -1, 0$, the treatment indicator $D(\omega) = 0$:

$$D(\omega) = \begin{cases} 0 & \text{if} \quad G(\omega) = -1, 0 \\ 1 & \text{if} \quad G(\omega) = 1 \end{cases}$$

Unconfoundedness only requires that $Y(\omega,0), Y(\omega,1) \perp\!\!\!\perp D(\omega)|X(\omega)$, but it is not testable. The focus is indeed on an implication of the stronger conditional independence relation $Y(\omega,0), Y(\omega,1) \perp\!\!\!\perp G(\omega)|X(\omega)$. This independence condition implies the unconfoundedness and also implies testable conditions:

$$Y(\omega,0) \perp\!\!\!\perp G(\omega)|X(\omega), G(\omega) \in \{-1,0\} \Longleftrightarrow Y(\omega) \perp\!\!\!\perp G(\omega)|X(\omega), G(\omega) \in \{-1,0\}$$

Because this new condition is stronger than the unconfoundedness, the question is whether there are interesting settings where the weaker condition holds, but not the stronger condition. To discuss this question, it is useful to consider two alternative conditions, both of which are implied by $Y(\omega,0), Y(\omega,1) \perp\!\!\!\perp G(\omega)|X(\omega)$:

$$Y(\omega,0), Y(\omega,0) \perp\!\!\!\perp D(\omega)|X(\omega), G(\omega) \in \{-1,1\}$$

and

$$Y(\omega,0), Y(\omega,0) \perp\!\!\!\perp D(\omega)|X(\omega), G(\omega) \in \{0,1\}\,.$$

For example, if the first condition holds the estimate of the average casual effect is done by invoking the unconfoundedness assumption using only the first control group. It is difficult to envision a situation where unconfoundedness based on the two comparison group holds ($Y(\omega,0), Y(\omega,1) \perp\!\!\!\perp D(\omega)|X(\omega)$), but it does not hold using only one of the two comparison groups at the time. In practice, it seems likely that if unconfoundedness holds then so will the stronger condition. The test consists in testing whether there is a difference in average values of $Y(\omega)$ between the two control groups, after adjusting for differences in $X(\omega)$:

$$\mathbb{E}\left[\mathbb{E}\left[Y(\omega)|G(\omega) = -1, X(\omega)\right] - \mathbb{E}\left[Y(\omega)|G(\omega) = 0, X(\omega)\right]\right] = 0.$$

A second set of tests of unconfoundedness focuses on estimating the casual effect of the treatment on a variable known to be unaffected by it, typically because its value is determined prior to the treatment itself. Such variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome. If it is not zero, this implies that the treated observations are distinct from the controls. If the treatment is instead zero, it is more plausible that the unconfoundedness assumption holds. Nevertheless if the variables used in the proxy test are closely related to the outcome of interest, the test arguably has more power. First partition the vector of covariates $X(\omega)$ into two parts, a pseudo outcome, denoted by $X^p(\omega)$, and the remainder, denoted by $X(\omega)^r$, so that $X(\omega) = (X(\omega)^p, X(\omega)^r)$. It is possible to consider the following modified unconfoundedness condition, which requires conditioning only on the subset of covariates $X(\omega)^r$:

$$Y(\omega,0), Y(\omega,0) \perp\!\!\!\perp D(\omega)|X(\omega)^r.$$

This is a stronger condition, indeed it is theoretically possible that conditional on a subset of the covariates unconfoundedness holds, but at the same time unconfoundedness does not hold conditional on the full set of covariates. In practice this situation is rare. Generally making subpopulation more homogeneous in pretreatment variables tends to improve the plausibility of unconfoundedness. The modified unconfoundedness condition is not testable, for the same reasons the original unconfoundedness assumption is not testable. Nevertheless, if one has a proxy for either of the potential outcomes, one can test independence for the proxy variable. In such case it is possible to use as proxy variable the pseudo outcome $X(\omega)^p$. That is, we view as a proxy for, say, $Y(\omega, 0)$, and asses the modified unconfoundedness by testing:

$$X(\omega)^p \perp\!\!\!\perp D(\omega) | X(\omega)^r.$$

An example is where $X(\omega)$ contains multiple lagged measures of the outcome. In such case, one can plausibly assess unconfoundedness by testing whether:

$$Y_{i,-1} \perp\!\!\!\perp D(\omega) | Y_{i,-2}, \ldots, Y_{i,-6}$$

where $Y_{i,-2}, \ldots, Y_{i,-6}$ are the lagged outcomes.

1.3.3 Methods based on propensity score: matching

A method widely used in literature is based on the propensity scores. Rosenbaum and Rubin (1983) show that under unconfoundedness, independence of potential outcomes and treatment indicators also holds after conditioning solely on the propensity score $e(x) = Pr(D(\omega) = 1|X(\omega))$:

$$D(\omega) \perp\!\!\!\perp (Y(\omega,0), Y(\omega,1)) \,|X(\omega) \Longrightarrow D(\omega) \perp\!\!\!\perp (Y(\omega,0), Y(\omega,1)) \,|e(X(\omega)).$$

For any binary variable $D(\omega)$, and any random vector $X(\omega)$, it is true that:

$$D(\omega) \perp\!\!\!\perp X(\omega) \,|e(X(\omega)).$$

Since under unconfoundedness all biases can be removed by adjusting for difference covariates, this means that within subpopulation homogeneous in the propensity score, covariates are independent of the treatment indicator and there are no biases in comparing treated and control units. This result can be exploited in a number of ways. Principally three of these have been used in practice.

The first method simply uses the *propensity score in place of the covariates* in regression analysis. Define :

$$v_d(e) = \mathbb{E}[Y(\omega)|D(\omega) = d, e(X(\omega)) = e].$$

Unconfoundedness in combination with the Rosembaum and Rubin result implies that $v_d(e) = \mathbb{E}[Y(\omega, d)|e(X(\omega)) = e]$. It is possible, then, to estimate $v_d(e)$ using kernel on propensity score, something which is greatly simplified by the fact that propensity score is a scalar.

The second method is labeled blocking, subclassifications or *stratification*. The idea is to partition the sample into strata by values of the propensity score, and then analyze the data with each stratum as if the propensity score were constant and the data could be interpreted as coming from a completely randomized experiment.

The third method exploiting the propensity score is based on *weighting*. Recall that $\tau_{ate} = \mathbb{E}[Y(\omega, 1)] - \mathbb{E}[Y(\omega, 0)]$ and consider the two terms separately. It is possible to demonstrate that weighting the treated population by the inverse of the propensity score recovers the expectation of the unconditional response under treatment:

$$\begin{aligned}
\mathbb{E}\left[\frac{D(\omega)\cdot Y(\omega)}{e(X(\omega))}\right] = \mathbb{E}\left[\frac{D(\omega)\cdot Y(\omega,1)}{e(X(\omega))}\right] = \mathbb{E}\left[\mathbb{E}\left[\left.\frac{D(\omega)\cdot Y(\omega,1)}{e(X(\omega))}\right| X(\omega)\right]\right] = \\
= \mathbb{E}\left[\frac{\mathbb{E}(D(\omega)|X(\omega))\cdot\mathbb{E}(Y(\omega,1)|X(\omega))}{e(X(\omega))}\right] = \\
= \mathbb{E}\left[\frac{e(X(\omega))\cdot\mathbb{E}(Y(\omega,1)|X(\omega))}{e(X(\omega))}\right] = \\
= \mathbb{E}\left[\mathbb{E}(Y(\omega,1)|X(\omega))\right] = \mathbb{E}\left[Y(\omega,1)\right].
\end{aligned}$$

The seqnarray* final inequalities follow by iterated expectations and the third equality holds by unconfoundedness. A similar calculation shows $\mathbb{E}\left[((1 - D(\omega))Y(\omega))/(1 - e(X(\omega)))\right] = \mathbb{E}\left[Y(\omega, 0)\right]$. Then:

$$\tau_{ate} = \mathbb{E}[Y(\omega,1)] - \mathbb{E}[Y(\omega,0)] = \mathbb{E}\left[\frac{D(\omega)\cdot Y(\omega)}{e(X(\omega))} - \frac{(1-D(\omega))Y(\omega)}{(1-e(X(\omega))}\right].$$

This result suggests the following estimator for τ_{ate}:

$$\hat{\tau}_{weight} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D(\omega)\cdot Y(\omega)}{e(X(\omega))} - \frac{(1-D(\omega))Y(\omega)}{1-e(X(\omega))}\right].$$

This estimator, as a sample average from a random sample, is consistent for τ_{ate} and is $\sqrt{N}$ asymptotically normally distributed. In practice, this is not a feasible estimator because it depends on the propensity score function which is rarely known. A surprising result is that, even if we know the propensity score, the estimator does not achieve the efficiency bound. Estimating the propensity score it is possible to achieve a more efficient estimator, asymptotically, than to know the propensity score. Replacing $e(\cdot)$ with a logistic

sieve estimator it is possible to construct the inverse probability weighting (IPW) estimator:

$$\hat{\tau}_{ipw} = \frac{\sum_{i=1}^{N} \frac{D(\omega) \cdot Y(\omega)}{\hat{e}(X(\omega))}}{\sum_{i=1}^{N} \frac{D(\omega)}{\hat{e}(X(\omega))}} - \frac{\sum_{i=1}^{N} \frac{(1 - D(\omega)) \cdot Y(\omega)}{1 - \hat{e}(X(\omega))}}{\sum_{i=1}^{N} \frac{D(\omega)}{1 - \hat{e}(X(\omega))}}$$

The blocking estimator can also be interpreted as a weighted estimator. A particular concern with IPW estimators arise when covariates distributions are substantially different from the two treatments groups, because some values of the propensity score get close to zero or one. That arises the problem of the model's choice because alternative parametric models are more different when the probabilities are so extreme. Moreover for units with propensity score close to zero and one, the weights can be large, making those units particularly influential in the estimates of the average treatment effect, and thus making the estimator imprecise. This problem can be less severe for the ATT parameter because propensity score values close to zero play no role.

One of the most used methods that exploit the propensity score is *Matching*. Matching estimators impute the missing potential outcomes using only the outcomes of a few nearest neighbors of the opposite treatment group. In that sense, matching is similar to non parametric kernel regression, with the number of neighbors playing the role of the bandwidth in the kernel regression (Figure 1.5). The difference between these two methods is that the asymptotic distribution for matching estimators is derived conditional on the implicit bandwidth, that is the number of neighbors, often fixed at a small number. The implicit estimate $\hat{\mu}_d(x)$ is unbiased, but not consistent, in contrast to the kernel estimators. Matching estimators have the two main attractive features: the smoothing parameters are easily interpretable and these estimators are easier to use than those estimators that require more complex choices of smoothing parameters. This estimators have been applied in setting where the interest is in the average treatment effect for the treated and where there is a large reservoir of potential controls. The setting with many potential controls allows to match each treated unit to one or more distinct controls, hence the label "matching without replacement". Given the matched pairs, the treatment effect within a pair is estimated as the difference in outcomes, and the overall average as the average of the within pair difference. Often to match units are used algorithms that sequentially match units. Most commonly the units are ordered by the value of the propensity score with the highest propensity score units matched first. Formally, given a sample $(Y(\omega), X(\omega), D(\omega))_{\omega=1}^{N}$, let $\mathscr{L}_m(\omega)$ be the index

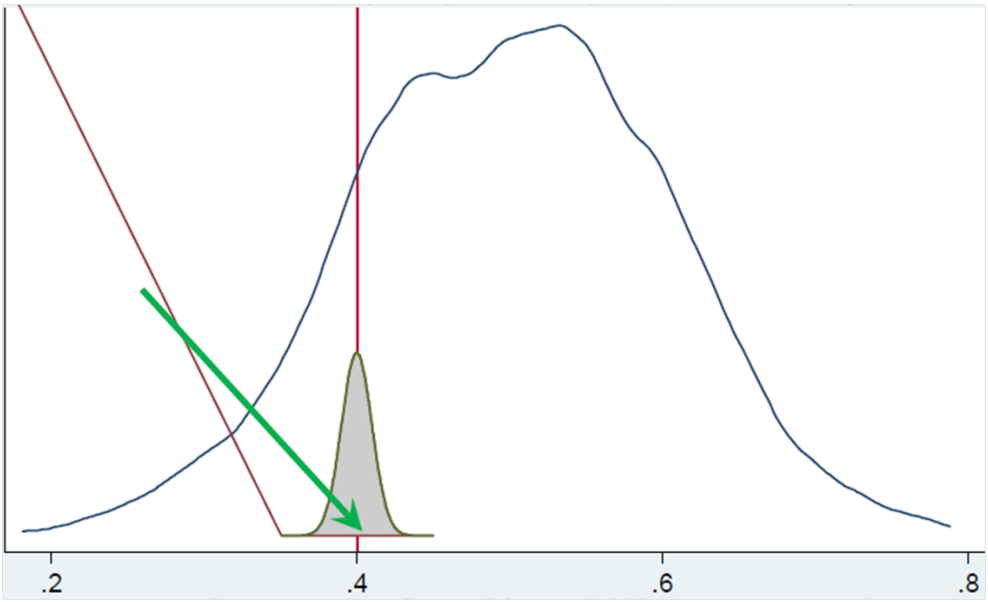

Figure 1.5: Example of propensity score distribution and kernel estimation

satisfies $D_{\mathscr{L}_m(\omega)} \neq D(\omega)$. This is the m-th closest to unit i:

$$\sum_{\omega' : D_{\omega'} \neq D_\omega} \mathbb{1}\left\{ \|X_l - X(\omega)\| \leq \left\|X_{\mathscr{L}_m(\omega)} - X(\omega)\right\| \right\} = m$$

where $\mathbb{1}\{\cdot\}$ is the indicator function which equals 1 when the event inside parenthesis is true and is zero otherwise. Definitely $\mathscr{L}_m(\omega)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit ω in term of distance measure based on the norm $\|\cdot\|$.

Let $\mathcal{J}_M(\omega) \subset 1, \ldots, N$ denote the set of index for the M matches for unit $\omega : \mathcal{J}_M(\omega) = \mathscr{L}_1(\omega), \mathscr{L}_2(\omega), \ldots, \mathscr{L}_M(\omega)$. Now impute the missing potential outcomes as the average of the outcomes for the matches, by defining $\hat{Y}(\omega, 0)$ and $\hat{Y}(\omega, 1)$ as

$$\hat{Y}(\omega, 0) = \begin{cases} Y(\omega) & \text{if} \quad D(\omega) = 0 \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(\omega)} Y_j & \text{if} \quad D(\omega) = 1 \end{cases}$$

$$\hat{Y}(\omega, 1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(\omega)} Y_j & \text{if} \quad D(\omega) = 0 \\ Y(\omega) & \text{if} \quad D(\omega) = 1 \end{cases}$$

The simple matching estimator discussed in Abadie and Imbens is then:

$$\hat{\tau}_{match} = \frac{1}{N} \sum_{\omega=1}^{N} \left(\hat{Y}(\omega, 1) - \hat{Y}(\omega, 0) \right).$$

Abadie and Imbens show that the bias of the estimator is of order $O(N^{-\frac{1}{k}})$, where K is the dimension of the covariates. Moreover they show also that matching estimators are generally not efficient.

1.4 Selection on unobservables

As described in Section 1.2.4 the assumption of unconfoundedness can be relaxed or dropped. Often in practice it is difficult to observe all the variables that can affect potential outcomes. The unconfoundedness condition is often unrealistic, given the presence of some unobservable or unmeasurable factors that affect the outcome. In literature there are different methods to solve the problem and to take into account the unobservable variables. The main approaches are listed below and the ones most used in practice are then summarized in the next paragraphs.

One approach (Rosenbaum and Rubin, 1983; Rosenbaum, 1995) consists of sensitivity analyses. Instead of completely relaxing the unconfoundedness assumption, sensitivity analyses investigate whether results obtained maintaining this assumption can be changed substantially, or even overturned entirely, by modest violations of it. Violations of unconfoundedness assumption are interpreted as evidence of the presence of unobserved covariates that are correlated, both with the potential outcome and with the treatment indicator. The size of bias this violations of unconfoundedness can induce depends on the strength of these correlations.

A second approach, developed by Manski (Manski, 1990, 1995, 2003, 2005, 2007) simply drops the unconfoundedness assumption. It consists of bound analyses, where ranges of estimate consistent with the data and the limited assumptions the researcher is willing to make, are derived and estimated. Manski's key insight is that even if in large samples one cannot infer the exact value of the parameter, one may be able to rule out some values that one could not *a priori*.

A third approach, instrumental variables, relies on the presence of additional treatments, the so-called instruments, that satisfy specific exogeneity and exclusion restrictions (Imbens and Angrist, 1994; Angrist *et al.*, 1996).

A fourth approach applies where overlap is completely absent because the assignment is a deterministic function of covariates. Comparisons can be made exploiting continuity of average outcomes as a function of covariates. This approach is known in statistics as the regression discontinuity design (Shadish *et al.*, 2002; Cook, 2008; Hahn *et al.*, 2001; Angrist *et al.*, 1996), but has recently been revived in the economics literature (Lee, 2001; Van der Klaauw, 2008; Imbens and Lemieux, 2008).

A fifth approach, referred to as difference-in-difference (DID), relies on the presence of additional data in the form of samples of treated and control units before and after the treatment (Ashenfelter and Card, 1985; Abadie, 2005; Donald and Lang, 2007; Athey and Imbens, 2006).

In the following paragraphs the most used methods are summarized: instrumental variable, regression discontinuity design and difference in difference (DID) methods.

1.4.1 Instrumental variables

The focus will be in the part of the literature concerned with heterogeneous effect and binary endogenous variable. Let $Z(\omega)$ denote the value of the instrument for individual ω. Let $D(\omega,0)$ and $D(\omega,1)$ denote the level of the treatment received if the instrument takes values 0 and 1 respectively. Similarly, let $Y(\omega,0)$ and $Y(\omega,1)$ denote the potential outcome of interest. The observed treatment is:

$$D(\omega) = D(\omega,0)\cdot(1-Z(\omega)) + D(\omega,1)\cdot Z(\omega) = \begin{cases} D(\omega,0) & \text{if} \quad Z(\omega)=0 \\ D(\omega,1) & \text{if} \quad Z(\omega)=1. \end{cases}$$

Exogeneity of the instrument is captured by the assumption that all potential outcomes are independent of the instrument:

$$(Y(\omega,0), Y(\omega,1), D(\omega,0), D(\omega,1)) \perp\!\!\!\perp Z(\omega).$$

This assumption captures two properties of the instrument. First, it captures random assignment of the instrument so that casual effects of the instrument on the outcome and treatment received can be estimated consistently. The second part of the assumption captures the exclusion restriction that there is no direct effect of the instrument on the outcome.

Imbens and Angrist introduce a new concept, the compliance type of an individual. The type of an individual describes the level of the treatment that an individual would receive given each value of the instrument. It is captured by the pair of values $(D(\omega,0), D(\omega,1))$. With both the treatment and instrument binary there are four types of responses for the potential treatment:

$$T(\omega) = \begin{cases} \text{never - taker} & \text{if } D(\omega,0) = D(\omega,1) = 0 \\ \text{complier} & \text{if } D(\omega,0) = 0, D(\omega,1) = 1 \\ \text{defier} & \text{if } D(\omega,0) = 1, D(\omega,1) = 0 \\ \text{always - taker} & \text{if } D(\omega,0) = D(\omega,1) = 1. \end{cases}$$

The labels never-taker, complier, defier and always-taker (Figure 1.6) refer to the setting of a randomized experiment with noncompliance, where the instrument is the random assignment to the treatment and the endogenous regressor is an indicator for the actual receipt of the treatment. Compliers are individuals who always comply with their treatment, that is, take the treatment if assigned to it and not to take it if assigned to the control group. One cannot infer from the observed data whether a particular individual is a complier or not. It is important not to confuse compliers with individuals who are observed to comply with the assignment they actually received $Z(\omega) = D(\omega)$. For such individuals it is unknown what they would have

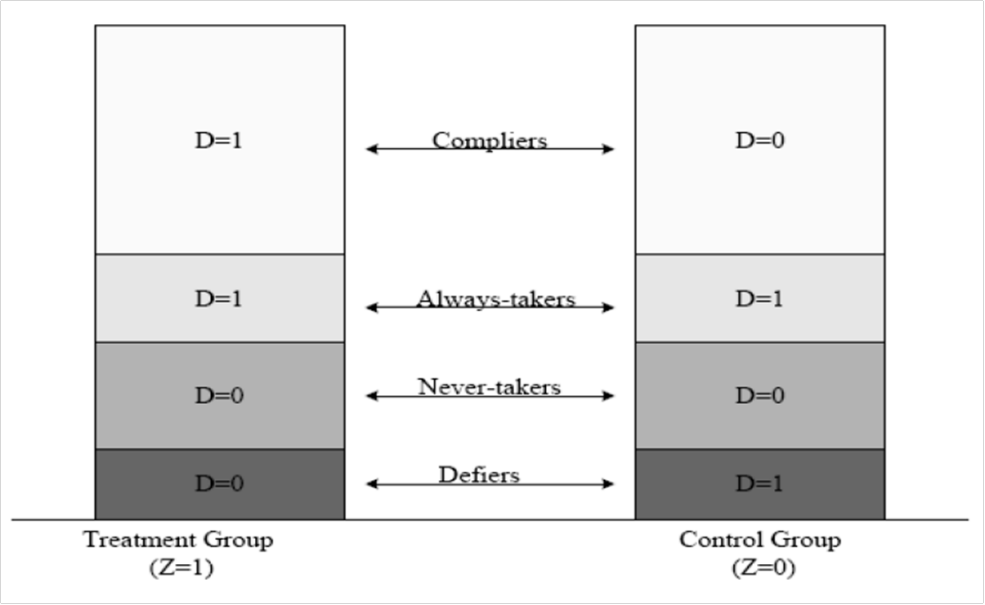

Figure 1.6: Instrumental variables: never-taker, complier, defier and always-taker

done had their assignment been different, that is it is unknown the value of $D(\omega)(1 - Z(\omega))$.

Imbens and Angrist then invoke an additional assumption they refer to *monotonicity*. It requires that $D(\omega, 1) \geq D(\omega, 0)$ for all individuals, or that increasing the level of the instrument does not decrease the level of the treatment. This assumption is equivalent to ruling out the presence of defiers, and it is therefore sometimes referred to as the no-defiance assumption.

Under the two assumptions - independence of all four potential outcomes and the instrument, and monotonicity - Imbens and Agrist show that one can identify the average effect of the treatment for the subpopulation of compliers. Thus, only compliers are observed in both treatment groups, so only for this group there is the chance of identifying the average treatment effect. Clearly, one cannot identify the average effect of the treatment for never-takers because they are never observed receiving the treatment. Individuals with $(Z(\omega) = 1, D(\omega) = 0)$ can only be never-takers for the monotonicity assumption. Similarly, individuals with $(Z(\omega) = 0, D(\omega) = 1)$ can only be always-takers. Individuals with $(Z(\omega) = 0, D(\omega) = 0)$ can be either compliers or never-takers, so it is not possible to infer the type of such individuals from the observed data alone. Similarly, individuals with $(Z(\omega) = 1, D(\omega) = 1)$ can be either compliers or always-takers.

The first step is to see if it is possible to infer the population proportions of these three subpopulation shares $P_t = P(T(\omega) = t)$, for $t \in \{n, a, c\}$. Considering the subpopulation of $Z(\omega) = 0$. Within this subpopulation it

is observed $D(\omega) = 1$ only for always-takers. Hence the conditional probability of $D(\omega) = 1|Z(\omega) = 0$ is equal to the conditional probability share of always-takers $P_a = Pr(D(\omega) = 1|Z(\omega) = 0)$. Similarly, for never-takers $P_n = Pr(D(\omega) = 0|Z(\omega) = 1)$. The population share of compliers is then obtained by subtracting the population shares of never-takers and always-takers from one $P_c = 1 - P_n - P_a$.

The second step uses the distribution of $Y(\omega)|(Z(\omega), D(\omega))$. The distribution of $Y(\omega)|D(\omega) = 0, T(\omega) = n$ is inferred from the subpopulation with $(Z(\omega), D(\omega)) = (1, 0)$ since all these individuals are known to be never-takers. Then the distribution of $Y(\omega)|Z(\omega) = 0, D(\omega) = 0$ is a mixture of the distribution of $Y(\omega)|D(\omega) = 0, T(\omega) = n$ and $Y(\omega)|D(\omega) = 0, T(\omega) = c$, with mixture probabilities equal to the relative population shares, $P_n/(P_c + P_n)$ and $P_c/(P_c + P_n)$, respectively. The conditional distribution of $Y(\omega)|D(\omega) = 0, T(\omega) = c$ is backed out from the population shares of the never-takers and compliers and the distribution of $Y(\omega)|D(\omega) = 0, T(\omega) = n$. Similarly, it is possible to infer the conditional distribution of $Y(\omega)|D(\omega) = 1, T(\omega) = c$. The difference between the means of these conditional distribution is the Local Average Treatment Effect (LATE):

$$\tau_{late} = \mathbb{E}[Y(\omega, 1) - Y(\omega, 0)|D(\omega, 0) = 0, D(\omega, 1) = 1] =$$

$$= \mathbb{E}[Y(\omega, 1) - Y(\omega, 0)|T(\omega) = \text{complier}].$$

Imbens and Angrist show that LATE equals the standard instrumental variables estimand, the ratio of the covariance of $Y(\omega)$ and $Z(\omega)$ and the covariance of $D(\omega)$ and $Z(\omega)$:

$$\tau_{late} = \frac{\mathbb{E}[Y(\omega)|Z(\omega) = 1] - \mathbb{E}[Y(\omega)|Z(\omega) = 0]}{\mathbb{E}[D(\omega)|Z(\omega) = 1] - \mathbb{E}[D(\omega)|Z(\omega) = 0]} =$$

$$= \frac{\mathbb{E}[Y(\omega)(Z(\omega) - \mathbb{E}[Z(\omega)])]}{\mathbb{E}[D(\omega)(Z(\omega) - \mathbb{E}[Z(\omega)])]},$$

which can be estimated using two-stage-least-squares. The only quantities not consistently estimable are the average effects for never-takers and always-takers.

1.4.2 Regression discontinuity design

The basic idea behind the Regression Discontinuity design (RD design) is that the assignment to the treatment is determined, either completely or partly, by the value of a predictor (the forcing variable $X(\omega)$) being on either side of a common threshold. This generated a discontinuity, sometimes of size one, in the conditional probability of receiving the treatment as a function of this particular predictor. The forcing variable is often itself associated with the potential outcome, but this association is assumed to be

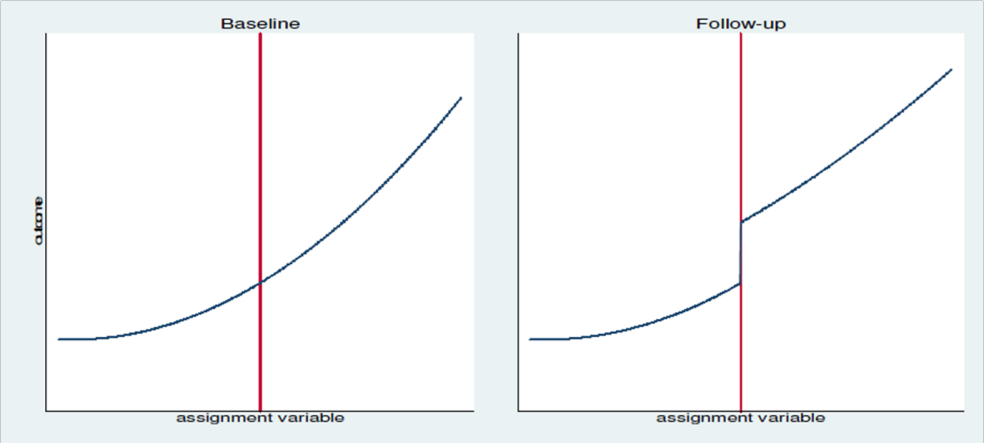

Figure 1.7: Discontinuity design: without and with treatment effect

smooth. As a result any discontinuity of the conditional distribution of the outcome as a function of this covariate at the threshold is interpreted as evidence of a casual effect of the treatment. This design often arises from administrative decisions, where clear transparent rules are used for the allocation of incentives for individuals to participate in a program. In the Regression Discontinuity (RD) design , the assignment $D(\omega)$ is a deterministic function of one of the covariates, the forcing variable $X(\omega)$:

$$D(\omega) = \mathbb{1}[X(\omega) \geq c],$$

where $\mathbb{1}[\cdot]$ is the indicator function. All units with a covariate value of at least c are in the treatment group, and all units with a covariate value less then c are in the control group. In the RD design the estimation is focused on:

$$\tau_{srd} = \mathbb{E}[Y(\omega,1) - Y(\omega,0)|X(\omega) = c] =$$

$$= \mathbb{E}[Y(\omega,1)|X(\omega) = c] - \mathbb{E}[Y(\omega,0)|X(\omega) = c].$$

By design there are no units with $X(\omega) = c$ for which it is observed $Y(\omega,0)$. To estimate $\mathbb{E}[Y(\omega)(w)|X(\omega) = c]$ without making functional form assumptions, it is exploited the possibility of observing units with covariate value arbitrary close to c. In order to justify this averaging it is made a smoothness assumption that the two conditional expectations are continuous in x. Under this assumption it is possible to write:

$$\tau_{srd} = lim_{x\downarrow c}\mathbb{E}[Y(\omega)|X(\omega) = x] - \mathbb{E}[Y(\omega,1)|X(\omega) = c] +$$

$$- lim_{x\uparrow c}\mathbb{E}[Y(\omega)|X(\omega) = x],$$

where this expression uses the fact that $D(\omega)$ is a deterministic function of $X(\omega)$. The statistical problem becomes one of estimating a regression function non parametrically at a boundary point.

1.4.3 Difference-in-difference method

The simplest setting is one where outcomes are observed for units observed in one of two groups, in one of the two time periods. Only units in one of the two groups, in the second time period, are exposed to the treatment. There are no units exposed to the treatment in the first period. The average gain over time in the non-exposed (control) group is subtracted from the gain over time in the exposed (treatment) group. This double differencing removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of time trends unrelated to the treatment.

The standard model for DID approach is that individual ω belongs to a group, $G(\omega) \in \{0,1\}$ is observed in time period $T(\omega) \in \{0,1\}$. The individual ω's group identity and time period can be treated as random variables. The outcome for the individual ω in the absence of the intervention is:

$$Y(\omega,0) = \alpha + \beta \cdot T(\omega) + \gamma \cdot G(\omega) + \epsilon(\omega),$$

with unknown parameter α, β and γ. The parameter β represents the time component to both groups; γ represents a group specific, time-invariant component; $\epsilon(\omega)$ represents unobservable characteristics of the individual. This term is assumed to be independent of the group indicator and have the same distribution over time $\epsilon(\omega) \perp\!\!\!\perp (G(\omega), T(\omega))$ and is normalized to have mean zero. The equation for the outcome without the treatment is combined with an equation for the outcome given the treatment: $Y(\omega,1) = Y(\omega,1) + \tau_{did}$. The standard DID estimand is under this model equal:

$$\begin{aligned}
\tau_{DID} &= \mathbb{E}[Y(\omega,1)] - \mathbb{E}[Y(\omega,0)] = \\
&= (\mathbb{E}[Y(\omega)|G(\omega)=1, T(\omega)=1] - \mathbb{E}[Y(\omega)|G(\omega)=1, T(\omega)=0]) + \\
&\quad - (\mathbb{E}[Y(\omega)|G(\omega)=0, T(\omega)=1] - \mathbb{E}[Y(\omega)|G(\omega)=0, T(\omega)=0]).
\end{aligned}$$

It is possible to estimate τ_{did} using least squares methods on the regression function for the observed outcome:

$$Y(\omega) = \alpha + \beta_1 \cdot T(\omega) + \gamma_1 \cdot G(\omega) + \tau_{did} \cdot D(\omega) + \epsilon(\omega),$$

where the treatment indicator $D(\omega)$ is equal to the interaction of the group and time indicators, $I(\omega) = T(\omega) \cdot G(\omega)$. The treatment effect is estimated

through the coefficient on the interaction between the indicators for the second time period and the treatment group:

$$\hat{\tau}_{did} = (\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00}) \quad \text{whith} \quad \overline{Y}_{gt} = \sum_{\omega|G(\omega)=g,T(\omega)=t} Y(\omega)/N_{gt}$$

where $\overline{Y}_{gt}$ is the average outcome among units in group g and the time period t.

Most of the recent econometric program evaluation literature has focused on the case with a binary treatment. Much less is known about setting with multi-valued, discrete or continuous treatments, that are common on practice. For more details see Wooldridge and Imbens (2008).

Chapter 2

The econometric approach to causality

The econometric structural approach is another class of methods which have been developed during the past decades to face the evaluation problem. The formulation of such problem in economics dates back to Quandt's work (1958). In this chapter the main features of this approach are presented with a comparison of the statistical approach described in the previous chapter. The main difference compared to the statistical approach, is the use of models to generate the counterfactual distribution. For this approach it is important to model the preferences and choices of the agents in order to infer both objective outcomes and subjective evaluation. A distinction between anticipated and realized subjective and objective outcomes is also developed. It also deals with the identification problem illustrating an example of a prototypical model of treatment choice and outcome. Finally, it reports a synthesis of the difference and of the points of contact between this approach and the potential outcome approach. It is a summary of many statistical and economical articles recently published aimed to read the connections between the two approaches. Among them there are also recent works by prime Nobel Heckman.

2.1 A model of hypotheticals

In the econometric tradition a model of the phenomena being studied has to be fully articulated to define hypothetical or counterfactual states. The definition of causality is a by-product of this model. Therefore ambiguity in model specification implies ambiguity of counterfactuals and, hence, in causality.

Opposition of economists toward statisticians starts by the fact that statisticians give a definition of causality without a clear model of the phenomenon studied, that is a model of counterfactuals. In economics a main

Task	Description	Requirement
1	Defining the set of Hypotheticals or counterfactuals	A well specified economic theory
2	Identifying causal parameters from Hypothetical population data	Mathematical analysis or Set identification
3	Identifying causal parameters from real data	Estimation and testing theory

Table 2.1: The three distinct tasks in the analysis of econometric causal models

point is to produce a clear model of how the studied phenomena are generated or what mechanism selects the counterfactuals that are observed in the real samples. Statisticians, on the contrary, want to model the effects of causes without modeling the causes of effects (Holland, 1986). However science is all about constructing models of the causes of effects. Heckman and Vytlacil (2007b) identify three distinct central tasks in economics:

1. defining the set of counterfactuals/hypotheticals;

2. identifying causal models from hypothetical data of population distribution (infinite samples without any sampling variation);

3. identifying causal model from real data, where sampling variability is an issue (it considered the difference between empirical distribution based on sampled data and population distribution generating the data).

The first task consists in defining counterfactuals. It is an essential step that arises from the need to represent the actual world trough models. These models are not a description of the actual world, but they are representations of empirical distribution that are used to make prediction about the actual world. These models are description of hypothetical worlds obtained by varying, hypothetically, the factors determining outcomes. The second task is the inference in very large samples. This is an identification problem that arises from the necessity to recover the distribution of counterfactual from data that are free of any sampling variation. The third task is the inference in practice. It is about recovering a given model of counterfactual from a given set of data, where solutions can be found in the inference and testing in real world data. The difference between the last two points arise from the difference between empirical distribution based on samples data and population distribution generating the data.

Some of the controversy surrounding counterfactuals and causal models is sometimes due to the confusion of these three distinct problems. Associating particular methods of estimation, like matching or instrumental

variables, with the definition of causal parameters, runs to confuse three distinct tasks: definition, identification and estimation.

2.1.1 Policy evaluation

The goal of the econometric literature is to understand the causes producing the effects. In this way one can use the model to forecast the effect of new policies never previously experienced, to calculate a variety of policy counterfactuals. Furthermore one can use scientific theory to choose estimators and interpret results.

In social science, a major use of causal analysis is directed toward answering policies questions. Heckman identifies three main classes of policy evaluation questions:

P1 evaluating the impact of historical (documented) interventions on outcome including their impact in terms of the well-being of the treated and society at large;

P2 forecasting the impacts (constructing the counterfactual states) of interventions implemented in one environment in other environment, including their impacts in terms of well-being;

P3 forecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced to various environments, including their impacts in terms of well-being.

In this context *impact* means constructing and evaluating either individual level or population level counterfactuals. *Well-being* means ex ante or ex post valuations of the outcomes obtained from the intervention.

Economists distinguish objective outcomes that can in principle be measured by all external observers from subjective outcomes that are the evaluations of the agents experiencing treatment. Objective outcomes are intrinsically ex post the treatment, while subjective outcomes can be ex ante (anticipated) or ex post. Agents may also have ex ante evaluations of the objective outcomes that may differ from their ex post evaluations.

The first problem, P1, is an *internal validity problem*, that is the problem of identifying a given treatment parameter or a set of treatment parameters in a given environment.

The second problem is helpful because most policy evaluation is designed with an eye toward the future and toward informing decisions about new policies and application of old policies to new environments, where the environment includes the characteristics of individuals and of the treatments. Included in these interventions are policies described by generic characteristics that are applied to different groups of people or in different time periods from those studied in implementations of the policies on which data

are available. This is the problem of *external validity*: taking a treatment parameter or a set of parameters estimated in one environment to another environment.

Finally, the most ambitious problem is forecasting the effect of a new policy, never experienced before. This problem requires one to use past history to forecast the consequences of new policies.

2.1.2 Individual level treatment effect: definition and notation

The individual ω does not represent a single individual, but encompass all individual's features that affect the outcomes. The universe of all individuals' types or agents is indicated with Ω and can be assumed to be $\Omega = [0, 1]$ (Heckman and Vytlacil, 2007a), with $\omega \in \Omega$.

The agent can be a household, a patient, a firm, a worker or a country. In advance of treatment, agents may not know the outcome but may make forecasts about them. These forecasts may influence their decisions to participate in the program or may influence the agents who make decisions about whether or not an individual participates in the program. Selection into the program based on actual or anticipated components of outcomes gives rise to the *selection problem* in the evaluation literature.

Let $\mathcal{S}$ be the set of possible treatments with elements denoted by s. For simplicity of exposition, let us assume that this set is the same for all ω and that it is finite. For each ω, one obtains a collection of possible outcomes given by $Y(s, \omega)_{s \in \mathcal{S}}$. For example, if $\mathcal{S} = 0, 1$, there are two treatments, one of which may be a no-treatment state (control).

The individual treatment effect for agent ω comparing outcomes of treatment s with outcomes of treatment s' is

$$Y(s, \omega) - Y(s', \omega), \quad s \neq s'$$

for two elements $s, s' \in \mathcal{S}$. This is also called an *individual level causal effect*. The causal effect is the Marshallian (1890) ceteris paribus change of outcomes for an agent across states s and s'. Only s and s' are varied. All factor save one (the treatment) are held at a constant level. In such way the change in the outcome is associated with the manipulation of the varied factor (the treatment that changes from s to s').

In econometrics, associated with each outcome, is also a valuation V of it. These valuations can be private evaluation of the agents with utility $V(Y(s, \omega), \omega)$, or may also be the valuation placed on the outcome of each person by another person, called the social planner, with preference V_G.

The valuation depends on which treatment is assigned to or chosen by the individual ω. The outcome is selected from the set of possible counterfactuals of potential outcomes available for each person.

The treatment assignment mechanism is a rule $\tau : \Omega \to \mathcal{S}$ which assigns treatment to each individual ω. The collection of the possible assignment rules is $\mathcal{T}$ where $\tau \in \mathcal{T}$ and the consequences of the treatment are the outcome $Y(s, \omega)$, $s \in \mathcal{S}$ and $\omega \in \Omega$. The policy selects individuals ω and specifies the treatment $s \in \mathcal{S}$ received.

Under a more comprehensive definition of treatment, agents are assigned incentives like taxes, subsidies and eligibility that affect their choices, but the agents choose the treatment selected. Agent's preferences, program delivery systems and the like might all affect the choice of treatment.

In a more general setup it is specified a rule a to assign such constraints or benefits b. The assignment rules $a \in \mathcal{A}$ maps individuals ω into constraints or benefits $b \in \mathcal{B}$ under different mechanisms $a : \Omega \to \mathcal{B}$ as a deterministic rule or random assignment. The last one adds a new source of randomness to the environment that it is necessary to consider redefining Ω to include it. Therefore it is possible to redefine the treatment assignment mechanism, that is the choice rule used by the agent, $\tau : \Omega \times \mathcal{A} \times \mathcal{B} \to \mathcal{S}$ as a map taking agent $\omega \in \Omega$ facing constraints $b \in \mathcal{B}$ assigned by mechanism $a \in \mathcal{A}$ into a treatment $s \in \mathcal{S}$. Moreover a policy regime $p \in \mathcal{P}$ is a pair $(a, \tau) \in \mathcal{A} \times \mathcal{T}$ that maps agents denoted by ω into elements of s.

2.1.3 Policy invariance

Policy invariance is used to characterize outcomes without specifying how those outcomes are obtained, ignoring features of the policy and choices environment.

Economists (see Heckman and Vytlacil (2007b)) define policy invariance through two aspects. The first aspect is that, for a given incentive $b \in \mathcal{B}$ the mechanism $a \in \mathcal{A}$ (by which ω is assigned a b) and the incentive $b \in \mathcal{B}$ are assumed to be irrelevant for the values of realized outcomes for each s that is selected. Second, for a given s for agent ω, the mechanism τ, by which s is assigned to the agent under assignment mechanism $a \in \mathcal{A}$, is irrelevant for the values assumed by realized outcomes. If one has to account for the effects of incentives and assignment mechanisms on outcomes, one must work with $Y(s, \omega, a, b, \tau)$ instead of $Y(s, \tau)$. These two aspects can be translated in the following invariance assumptions invoked in the literature.

The first invariance assumptions state that for the same treatment s and agent ω, different constraint assignment mechanisms a and a' and associated constraint state assignments b and b' produce the same outcome. For example, they rule out the possibility that the act of randomization or the act of pointing a gun at an agent to secure cooperation with planner intentions has an effect on outcomes, given that the agent ends up in s. This is a strong assumption.

The second invariance assumption is that for a fixed a and b, the outcomes are the same, independent of the treatment assignment mechanism.

It rules out, among other things, social interactions, contagion and general equilibrium effects. Heckman (1992), Heckman and Smith (1998), Heckman *et al.* (1999) and Heckman and Vytlacil (2007c) discuss evidence against this assumption, and Heckman *et al.* (1998a,b,c) show how to relax it.

If treatment effects based on subjective evaluations are also considered, as it is distinctive of the econometric approach, it is necessary to require invariance assumptions for utilities, that states that utilities are not affected by the mechanism of assignment of constraints. Heckman (1992), Heckman *et al.* (1999) and Heckman and Vytlacil (2007c) present evidence against this assumption.

Another invariance assumption rules out social interactions in both subjective and objective outcomes. It is useful to distinguish invariance of objective outcomes from invariance of subjective outcomes. Randomization may affect subjective evaluations through its effect of adding uncertainty into the decision process but it may not affect objective valuations. The econometric approach models how assignment mechanisms and social interactions affect choice and outcome equations rather than postulating a priori that invariance.

2.1.4 The evaluation problem

The *evaluation problem* is an *identification problem* that arises in constructing the counterfactual states and treatment assignment rules produced by abstract models from population data. This is the second econometric task identified in Table 2.1. As already underlined in Chapter 1, the analyst observes each agent ω in one of $\mathcal{S}$ possible states. Rarely the same person ω is observed in distinct state s. One does not know the outcome of the agent in other states that are not realized, and hence cannot directly form individual level treatment effects. The states are mutually exclusive. Let $D(s,\omega) = 1$ if we observe ω in state s. Then $D(s',\omega) = 0$ for $s \neq s'$. Thus the analyst cannot observe $Y(s',\omega)$ for person ω if he observes $Y(s,\omega)$, $s \neq s'$. Even with large sampled and a valid randomization, some of the $s \in \mathcal{S}$ may not be observed if one wants to evaluate new policies never experienced. Without further assumptions, constructing the counterfactual is impossible from the data $(Y(\omega), D(\omega))$, $\omega \in \Omega$. The formulation of the evaluation problem is known as *Quandt's switching regression model* (Quandt, 1972, 1958) in econometrics and is attributed in statistics to Neyman (1990) and Rubin (1978).

In addition to this problem, there is the *selection problem*: the values of $Y(0)$ or $Y(1)$ that are observed are not necessarily a random sample of the potential $Y(0)$ or $Y(1)$ distribution.

The Roy model (Roy, 1951) and his generalizations in economics (see Heckman and Smith (1998), Carneiro *et al.* (2003) and Cunha *et al.* (2007)) are a useful framework for policy evaluation. Roy considered an economy

where agents face two possible outcome $\mathcal{S} = (0, 1)$ and a particular selection mechanism:

$$D(1, \omega) = \mathbb{1}(Y(1, \omega) > Y(0, \omega)),$$

where "$\mathbb{1}$" is an indicator function. The mechanics of selection process depends on outcomes. Agents choose treatment with the expected highest outcome, so the selection mechanism is not a randomization. In this case there is a self-selection into treatment, so there is no independence between assignment rule and outcome. On one hand this selection rule creates the potential for selection bias. On the other hand the choice of treatment provides also information on subjective evaluations of treatment which are of independent interest in economics.

Statisticians attempts to create assignment rules so that $D(s, \omega)$ is random with respect to outcome $Y(s, \omega)$. This means that the receipt treatment is independent of the treatment outcome. They use different methods, like matching and instrumental variables, to create such situations.

The evaluation problem is solved from econometricians and statisticians in different ways. The econometric way consists in modelling $Y(s, \omega)$ explicitly in term of its determinant as specified by theory. This models are called *structural econometric analysis* and entails describing the random variables characterizing ω and carefully distinguish what agents know and what the analyst knows. This approach also models $D(s, \omega)$ and the dependence between $Y(s, \omega)$ and $D(s, \omega)$ produced from their variables in common. See Heckman and Honore (1990) and Heckman and Vytlacil (2001) for a discussion of the Roy model that models this dependence. The goal of the econometric literature, like the goal of all science, is to understand the causes producing effects so that one can use empirical versions of the models to forecast the effects of interventions never experienced before, to calculate a variety of policy counterfactuals and to use scientific theory to guide the choices of estimators and the interpretation of the evidence. These activities require development of a more elaborate theory than is envisioned in the current literature on causal inference in statistics.

Many causal models in statistics are designed to investigate the impact of treatments on observed outcomes in a given environment. Explicit scientific models try to explore the mechanisms producing the effects. In the terminology of Holland (1986), the distinction is between understanding the "effects of causes" (the goal of the treatment effect literature as a large group of statisticians define it) or understanding the "causes of effects" (the goal of the econometric literature building explicit models). For this reason statisticians call this approach the "scientific approach" and are hostile to it (see Holland (1986)).

A second way to solve the problem, common in the statistic literature, is to estimate some population version of the individual treatment effect,

often a mean, without modeling what factors give rise to the outcome or the relationship between the outcomes and the mechanism selecting outcomes. Following this avenue the agent valuations of outcomes are ignored. In fact, the statistical treatment effect literature focuses exclusively on policy problem P1 for the subset of outcomes that is observed. The treatment effect approach does not model the factors determining outcomes, it works solely with outcomes without inputs, differently from the econometric approach that uses data to generate counterfactual. In such a way they focus treatment effects for policies actually experienced and provide no framework for extrapolation of findings to new environments (problem P2) or for forecasting new policies (problem P3).

2.1.5 Population level treatment parameters

Economists and statisticians often use the same set of population level treatment parameters such as the average treatment effect (ATE), the treatment on the treated (TT) and the treatment on the untreated (TUT) explained in the previous chapter. In such way summary measures of outcomes are considered, not analyzing determinants of outcomes. Furthermore this approach, the only one used in statistics, confines attention to the subset of $\mathcal{S}$ that are observed states. Many other mean treatment parameters can be defined depending on the choice of the conditioning set. Analogous definitions can be given for median and other quantile versions of these parameters (see Heckman *et al.* (1997), Abadie *et al.* (2002). Although means are conventional, distributions of treatment parameters are also of considerable interest.

Economists use different causal parameters for different policy problems. It is of interest to evaluate the impact of marginal extensions (or contraction) of a program or treatment regime. For example, the cost-benefit analysis is conducted in terms of marginal gains and benefits. The *effect of treatment for people at the margin of indifference* (EOTM) is calculated between the best two possible choices available with respect to the personal preference and choice-specific cost. This represents the mean gain to people indifferently between the best two options available. A generalization of this parameter is the *Marginal Treatment Effect* developed in Heckman and Vytlacil (1999) Heckman and Vytlacil (2005) Heckman and Vytlacil (2007a).

Of special interest in policy analysis is the *policy relevant treatment effect* (PRTE). It is the effect on aggregate outcomes of one policy regime $p \in \mathcal{P}$ compared to the effect of another policy regime. Under invariance assumptions and with $p, p' \in \mathcal{P}$

$$\mathbb{E}_p[Y(s)] - \mathbb{E}_{p'}[Y(s)].$$

Usually the elevation of population means as the primary causal parameters promotes randomization as an ideal estimation method. Mean

treatment effects are easily identified, assuming full compliance and no bias arising from the randomization, thanks to the special mathematical property of means. If one can identify the mean of $Y(j)$ and the mean of $Y(k)$ from two different groups of agents, where j is the treatment and k is the baseline or the control, one can form the average treatment effect for j compared to k. The case for randomization is weaker if the analyst is interested in other summary measures of the distribution, as median, quantiles or the distribution itself. The answers to many interesting evaluation questions, in fact, require knowledge of other features of the distribution of program.

It is also of economic interest to know:

- the proportion of people taking the program j with benefits from it relative to some alternative k, that is $P_\omega(Y(j,\omega) > Y(k,\omega)|D(j,\omega) = 1)$. This measure is used in determining how program gains are distributed among participants;
- *voting criterion* : the proportion of the total population that benefits from the program k compared to program j, that is $P_\omega(Y(j,\omega) > Y(k,\omega))$. It measures the proportion of the entire population that benefits from a program;
- selected quantiles of the impact distribution. It reveals the gains at different percentiles of the impact distribution;
- the distribution of gains at selected base state values. This criterion focuses on the distribution of impacts for subgroups of participants with particular outcomes in the non participation state.

All of these measures require knowledge of features of the joint distribution of outcomes and not just the mean. Distribution of counterfactuals is necessary. Heckman and Smith (1998), Carneiro *et al.* (2003), Carneiro *et al.* (2001), Heckman and Navarro-Lozano (2004) Cunha *et al.* (2005) develop different methods for identifying it.

2.1.6 Different information sets: ex post and ex ante evaluation

Economic analysis account for uncertainty arising from the imperfect information of the agent. This uncertainty creates a distinction between *ex ante* and *ex post* evaluation of both subjective and objective outcomes. Ex ante and ex post distinction is essential to understanding behavior. Agent choices are made in term of ex ante calculations while effect literature usually reports ex post return. Let $\mathcal{I}_\omega$ denote the information set available to agent ω to evaluate policy j against k. Under an expected utility criterion, agent ω prefers policy j over policy k if

$$\mathbb{E}[R(Y(j,\omega),\omega)|\mathcal{I}_\omega] > \mathbb{E}[R(Y(k,\omega),\omega)|\mathcal{I}_\omega].$$

Ex post uncertainty arises because people do not know the outcome associated with possible states not experienced yet and/or do not know some outcomes. In advance of choosing an activity, agents may be uncertain about the outcomes that will actually occur. They may also be uncertain about the full costs they will bear. In general the agents' information is not the same as the analysts', and they may not be nested. The agent may know things in advance that the analyst may never discover. On the other hand, the analyst, benefiting from hindsight, may know some information that the agent does not know when he is making his choices. Let $\mathcal{I}_A$ be the information set confronting the agent at the time choices are made before outcomes are realized. Agents can only imperfectly estimate consequences of their choice. Ex ante agent's evaluation are based on determined variables that are known to the econometrician and other variables known to the agent but not to the econometrician. The econometrician may in fact posses different information set $\mathcal{I}_E$. Choice probabilities computed against one information set are not generally the same as those computed against another information set. Carneiro *et al.* (2001, 2003), Cunha *et al.* (2005, 2006) and Heckman and Navarro (2007) develop econometric methods for distinguishing ex ante from ex post evaluations of social programs. See Abbring and Heckman (2007) for an extensive survey of this literature. Heckman and Vytlacil (2007b) discuss the data needed to identify these criteria, and present examples of Roy models and their extensions that allow for more general decision rules and imperfect information by agents. They show how to use economic models to form treatment parameters. A prototypical econometric model for policy evaluation is the generalized Roy model developed in Cunha *et al.* (2005). A patient can be treated or untreated with outcomes $Y_1(\omega)$ and $Y_0(\omega)$ (I will drop the ω notation to simplify the notation) that is an index of well being of the patient. At any point in time, a person can be either treated or untreated. The decision to treat may be made on the basis of the expected outcomes $\mathbb{E}(Y_1|\mathcal{I})$ and $\mathbb{E}(Y_0|\mathcal{I})$ and costs $\mathbb{E}(C|\mathcal{I})$ where the expectations are those of the relevant decision maker. For example, if the agent is a hospital patient, then the costs might be the pain and suffering and/or the direct medical costs. From the point of view of the patient the expected utility or value of treatment is $\mathbb{E}(Y_1|\mathcal{I}) - \mathbb{E}(C|\mathcal{I})$. The value of no treatment is $\mathbb{E}(Y_0|\mathcal{I})$. The expected net value is $\mathbb{E}(Y_1|\mathcal{I}) - \mathbb{E}(C|\mathcal{I}) - \mathbb{E}(Y_0|\mathcal{I})$. For patients who pick a treatment based on maximum gain $D = \mathbb{1}[(\mathbb{E}(Y_1|\mathcal{I}) - \mathbb{E}(C|\mathcal{I}) - \mathbb{E}(Y_0|\mathcal{I})) \geq 0]$. The ex post treatment effect is $Y_1 - Y_0$ and the ex ante effect is $\mathbb{E}(Y_1|\mathcal{I}) - \mathbb{E}(Y_0|\mathcal{I})$. The econometric approach models the dependence between observed $Y = DY_1 + (1 - D)Y_0$ and D suggest alternative estimators to identify causal

parameters. Commonly used specifications (Heckman, 2008) are:

$$Y_1 = X\beta_1 + U_1$$

$$Y_0 = X\beta_0 + U_0$$

$$C = Z\gamma + U_C$$

where (X, Z) are observed by the analyst and (U_1, U_0, U_C) are unobserved.

The agent may know more or less than the analyst. Econometric models allow the agent to know more than the analyst and analyse agent selection into treatment accounting for the asymmetry in knowledge between them (matching assumes that, conditional on X and Z, D is independent of Y_0, Y_1 and so assumes a lot of information is available to the analyst).

The most basic Roy model assumes that decision makers information is perfect. There are no direct costs of treatment ($\gamma = 0$ and $U_C = 0$), the decision rule is $D = \mathbb{1}(Y_1 \geq Y_0)$ and assumes normality for (U_0, U_1). These distribution and parametric assumptions are relaxed in the recent econometric literature (see Heckman and Vytlacil (2007b) for a review).

The econometric approach makes the treatment assignment equation the centerpiece of its focus and considers both objective and subjective valuations as well as ex ante $(\mathbb{E}(Y_1|\mathcal{I}), \mathbb{E}(Y0|\mathcal{I}), \mathbb{E}(C|\mathcal{I}))$ and ex post outcomes (Y_1, Y_0, C).

2.1.7 Generating counterfactual

The traditional model of econometrics is the "all causes" model. It writes outcomes as a deterministic mapping of inputs to outputs:

$$y(s) = g_s(x, u_s) \tag{2.1}$$

where x and u_s are fixed variables specified, observable (x) and unobservable (u_s), by the relevant economic theory. In general, there is no objective way to choose these conditioning variables. Any argument for inclusion or exclusion of variables has to be made by an appeal (implicit or explicit) to theory. The role of the two types of variables is symmetric. This notation allows for different unobservables u_s to affect different outcomes. All outcomes are explained in a functional sense by the arguments of g_s, the equation maps admissible inputs into possible ex post outcomes. If one models the ex post realizations of outcomes, it is entirely reasonable to invoke an all causes model since the realizations are known (ex post) and all uncertainty has been resolved. Implicit in the definition of a function is the requirement that g_s be stable or invariant to changes in its arguments, x and u_s. A deep structural version of the equation above models the variation of the g_s in terms of s as a map constructed from generating characteristics q_s, x and

u_s into outcomes:

$$y(s) = g_s(q_s, x, u_s). \tag{2.2}$$

The components q_s provide the basis for generating the counterfactuals across treatments from a base set of characteristics. This framework provides the basis for solving policy problem P3 since new policies (or treatments) are generated from common characteristics, and all policies are put on a common basis. If a new policy is characterized by known transformations of (q_s, x, u_s) that lie in the domain of definition of g, policy forecasting problem P3 can be solved. The argument of the maps g_s and g are part of the a priori specification of a causal model, then analysts may disagree about appropriate arguments to include in these maps. In addition, modeling the u_s and its relationship with the corresponding unobservables in the treatment choice equation is highly informative on the choice of appropriate identification strategies.

These two models are sometimes called *Marshallian causal functions*. Assuming that the components of (x, u_s) or (q_s, x, u_s) are variation-free (can be independently varied coordinate by coordinate), one may vary each argument of these functions to get a ceteris paribus causal effect of the argument on the outcome.

Some components may be variation-free while others are not. These thought experiments are conducted for hypothetical variations. Varying q_s fixes different treatment levels. Variations in u_s among agents explain why people with the same x characteristics respond differently to the same treatment s. A treatment generally consists of a package of characteristics and if one varies the package from q_s to q_{s_0} one gets different treatment effects. Thus if uncertainty is a feature of the environment, these models can be interpreted as ex post realizations of the counterfactual as uncertainty is resolved.

Ex ante versions is represented with the ex ante expected value of $Y(s, \omega)$ conditioning on the information set of the agent $\mathcal{I}_A$

$$\mathbb{E}[Y(s, \omega)|\mathcal{I}_A] = \mathbb{E}[g_s(Q(s, \omega), X(\omega), U(s, \omega))|\mathcal{I}_A]$$

where $Q(s, \omega)$, $X(s, \omega)$, $U(s, \omega)$ are random variables generated from a distribution that depends on the agents information set indexed by $\mathcal{I}_A$. The expectation might be computed using the information set of the relevant decision maker (e.g. the parents in the case of the outcomes of the child) who might not be the agent whose outcomes are measured. These random variables are drawn from agents subjective distribution, that may differ from the distribution produced by reality if agent expectations are different from objective reality. In the presence of intrinsic uncertainty, the relevant decision maker acts on $\mathbb{E}[Y(s, \omega)|\mathcal{I}_A]$ but the ex post counterfactual is

$$Y(s, \omega) = \mathbb{E}[Y(s, \omega)|\mathcal{I}_A] + \nu(s, \omega) \tag{2.3}$$

where $\nu(s,\omega)$ satisfies $\mathbb{E}[\nu(s,\omega)|\mathcal{I}_A] = 0$. In this interpretation, the information set of agent ω is part of the model specification but the realizations come from a probability distribution, and the information set includes the technology g.

This representation clarifies the distinction between deterministic ex post outcomes and intrinsically random ex ante outcomes. Ex ante, there is uncertainty at the agent level but ex post there is not. The realizations of $\nu(s,\omega)$ are ingredients of the ex post "all causes" model, but not part of the subjective ex ante "all causes" model.

The ex ante treatment effect from the point of view of the agent for treatment s and s' is

$$E[Y(s,\omega|\mathcal{I}_A)] - E[Y(s',\omega|\mathcal{I}_A)]$$

Abbring and Heckman (2007) survey econometric evaluation models accounting for uncertainty.

The value of structural approach to the construction of counterfactuals is that it explicitly models the unobservables and the sources of variability among observationally identical people. It is the unobservable that gives rise to selection bias. Analyst can also use choice theory to model the choice of treatment to control for the selection bias.

These models derive from theory and the arguments of these functions can be hypothetically manipulated to produce outcome. It is necessary to be careful in distinguishing between theoretical relationship and empirical relationship.

Data used to determine these functions may be limited in their support. In such way it is not possible to identify the theoretical relationships. In addition, to this the component may not be variation-free even if they are in the hypothetical support. A good example is the problem of multicollinearity. If the X in the sample is linearly dependent, it is not possible to identify the Marshallian causal function with respect to variations in x over the available support even if one can imagine hypothetically varying the component of x over the domains of definition of the function.

In empirical data, one of the X, for example gender, may be perfectly predictable by the other X. Holland (1988) claims that the causal effects of gender is meaningless because analysts cannot "in principle" randomly assign gender. The problem is that he conflates the empirical problem of estimating it with a problem of theory of defining it. One can define the effect even if one cannot identify it from the population sample data. For example, with the local average treatment effect "LATE" parameter of Imbens and Angrist (1994) the effect is defined by an instrument and conflates definition and identification. The instrumental variables are used as surrogates for randomization.

2.2 Identification problem

A problem that must be solved if causal model has to be empirically operational is the *identification problem*. The act of defining a model is distinct from the one of estimating it. The identification problem asks whether theoretical models have an empirical content in a hypothetical population or in a real sample (problem 2 and 3 in Table 2.1). It is necessary to consider what particular models within a set of admissible models, produced by some theory for generating counterfactual, are consistent with a given set of data. After identifying the object of interest, that could be only a feature of the model (i.e. the average treatment effect) it is important to identify the model consistent with the available information.

Estimators differ in the amount of knowledge they assume that the analyst has on agent's decision. If the analyst has no access to all of the relevant information that produces the dependence between outcome and treatment rules, then he must use method to control for that unobserved component. The dependence between outcome and choice is the source of selection bias. Heckman and Vytlacil (2005) and Heckman and Navarro-Lozano (2004) define *relevant information* as information which, if available to the analyst, would eliminate selection bias.

Common to all scientific models, there is the additional issue of how to select the conditioning variables and how to deal with them if they are endogenous. Furthermore there is the problem of lack of knowledge of functional forms of the models. Different economic methods solve these problems in different way. For a discussion of the problem identification see Heckman (2005).

2.2.1 A prototypical model of treatment choice and outcome

To focus the discussion a benchmark econometric model of treatment choice and treatment outcome is presented (see Heckman (2005)). For simplicity a binary outcome is considered Y_0, Y_1.

Let V be a function (μ_V) of observed W (by the econometrician) and unobserved U_V factors determining choice then the assignment mechanism can be written as:

$$D = \mathbb{1}(V > 0).$$

Let potential outcomes be functions (μ_0, μ_1) of observed (X) and unobserved outcome-specific variables (U_0, U_1). Assuming additive separability between

factors and finite means of unobserved factors, it is possible to write:

$$V = \mu_V(W, U_V) \quad \rightarrow \quad V = \mu_V(W) + U_V \quad \mathbb{E}(U_V) = 0$$

$$Y_0 = \mu_0(X, U_0) \quad \rightarrow \quad Y_0 = \mu_0(X) + U_0 \quad \mathbb{E}(U_0) = 0$$

$$Y_1 = \mu_1(X, U_1) \quad \rightarrow \quad Y_0 = \mu_1(X) + U_1 \quad \mathbb{E}(U_1) = 0$$

Supposing linearity in parameters the potential outcome equations can be rewritten as:

$$\begin{aligned} Y_0 = \mu_0(X) + U_0 \rightarrow \quad & X\beta_0(C_0) + U_0 \\ Y_1 = \mu_1(X) + U_1 \rightarrow \quad & X\beta_1(C_1) + U_1 \end{aligned} \tag{2.4}$$

where X represents the characteristics of persons and β depends on C_1 and C_0,the characteristic of the program. These are linear in parameters versions of Equation (2.1). The U_1 and U_0 are the unobservables arising from omitted X, C_1 and C_0 components.

By modeling how β_1 and β_0 depend on C_1 and C_0 one can answer policy question P-3 for new programs that offer new packages of C. A version of the model most favorable to solving problem P-2 and P-3 is:

$$\beta_0(C_0) = \Lambda C_0'$$

$$\beta_1(C_1) = \Lambda C_1'$$

For each set of characteristics of a program one can model how outcomes are expected to differ when the characteristics of the people participating in them change (the X). Equations in (2.4) are in ex post all causes form. When the agent information is present one can tale into account agent's uncertainty about X, β_i, C_i and U_i using the ex ante version of outcomes $\mathbb{E}(Y_1|\mathcal{I})$ and $\mathbb{E}(Y_0|\mathcal{I})$. Ex ante functions are defined in terms of variation of $\mathcal{I}$ and are connected to ex post outcome by shock $\nu(s, \omega)$ as in Equation (2.3).

As explained in Section 2.1.6 agent chooses the treatment depending on subjective valuation of the outcome, given the own information set. Let V be the agent's valuation of treatment:

$$V = \mathbb{E}(Y_1 - Y_0 - (P_1 - P_0)|\mathcal{I})$$

where P_i is the price of participating in treatment i and $P_i = Z\varphi_i + \eta_i$. Leaving C_i implicit and making the right substitution it is obtained:

$$\begin{aligned} V &= \mathbb{E}(Y_1 - Y_0 - (P_1 - P_0)|\mathcal{I}) = \\ &= \mathbb{E}(X(\beta_1 - \beta_0) - Z(\varphi_1 - \varphi_0) + (U_1 + U_0) - (\eta_1 - \eta_0))|\mathcal{I}). \end{aligned}$$

Let $W = (X, Z)$ be the observed (by the econometrician)factors determining choice and $U_W = (U_1 - U_0) - (\eta_1 - \eta_0)$ the unobserved factors ones. Let $\gamma = ((\beta_1 - \beta_0) - (\varphi_1 - \varphi_0))$ be the vector of parameters, then the choice equation can be represented as:

$$V = \mathbb{E}(W_\gamma + U_W|\mathcal{I})$$

with $D = \mathbb{1}(V > 0)$ and where $U_V = U_W|\mathcal{I}$. Let assume that both agent and econometrician know W. The Roy model is a particular case with parameters $\varphi_0 = \varphi_1 = 0$, $\eta_0 = \eta_1 = 0$.

The selection problem arises when D is correlated with outcomes (Y_0, Y_1). This happens if the observables or unobservable in (Y_0, Y_1) are correlated with D. Thus there may be common observed or unobserved factors connection V and (Y_0, Y_1). If D is not independent of (Y_0, Y_1), the observed (Y_0, Y_1) are not randomly selected from the population distribution of (Y_0, Y_1). In particular, in the Roy model, selection is based on Y_0 and Y_1, indeed $D = \mathbb{1}(Y_1 > Y_0)$. Thus it is observed Y_1 if $Y_1 > Y_0$ and Y_0 if $Y_0 > Y_1$. If conditioning on W makes (Y_0, Y_1) independent of D, selection on observables is said to characterize the selection process (see Heckman and Robb Jr (1985)). If conditionals on W, (Y_0, Y_1) are not independent of D, then there is selection on unobservables. For the Roy model, Heckman and Honore (1990) show that it is possible to identify the distribution of treatment $(Y_1 - Y_0)$ under certain conditions. Randomization can only identify the marginal distribution of Y_1 and Y_0 and not the joint distribution of $(Y_1 - Y_0)$.

2.3 Comparison

Tukey (1986) underlines that the econometric approach to policy evaluation emphasizes the provisional nature of causal knowledge, because human knowledge advanced by developing theoretical models and testing them against data. The models are inevitable provisional and depend on a priori assumption. Even randomization cannot answer all of relevant causal questions.

Statisticians reject the notion of the provisional nature of causal knowledge and look for an assumption free approach to causal inference. They are motivated by the experiment as an ideal. They do not clearly specify the mechanism determining how hypothetical counterfactuals are realized. They do not model both the factors determining the outcome $Y(s, \omega)$, as econometrics do in Equations (2.1) and (2.2), and the choice of which outcome is selected. They focus only on outcomes, leaving the model for selection outcomes only implicitly specified. The emphasis on randomization or its surrogates, like matching or instrumental variables, rules out the identification of counterfactuals from population or sample data by formal models.

Since randomization is used to define the parameters of interest, this practice sometimes leads to the confusion that randomization is the only way, or the best way, to identify causal parameters. Unlike the NeymanRubin model, the econometric models do not start with the experiment as an ideal but they start with well-posed, clearly articulated models for outcomes and treatment choice where the unobservables that underlie the selection and evaluation problem are made explicit. The hypothetical manipulations define the causal parameters of the model. Randomization is a metaphor and not an ideal or "gold standard".

Econometricians say that statistical models are incomplete for various reasons. One reason is that they do not specify the sources of randomness generating variability among agents, so the source of variability generating $Y(s, \omega)$ as a random variable. In this way they do not specify why otherwise observationally identical people make different choices and have different outcomes given the same choice. Holland (1986, 1988) argues that it is an advantage of the Rubin model that these features are not explicit.

They do not distinguish what is in the agents' information set from what is in the observing statisticians' information set, although the distinction is fundamental in justifying the properties of any estimator for solving selection and evaluation problems. They do not distinguish uncertainty from the point of view of the agent whose behavior is being analyzed from variability as analyzed by the observing analyst.

They are also incomplete because they are recursive. They do not allow for simultaneity in choices of outcomes of treatment that are at the heart of game theory and models of social interactions. Since Haavelmo (1943), econometricians have used simultaneous equations theory to define causality in non-recursive models where causes are simultaneous and interdependent. Heckman (2005) and Heckman and Vytlacil (2007b) present extensive discussions of simultaneous causality.

The econometric framework is explicit about how models of counterfactuals are generated, the rules of assigning treatment, and the sources of unobservables in treatment allocations and outcomes and their relationship. Rather than leaving the rules governing selection of treatment implicit, the econometric approach uses explicit relationships between the unobservables in outcome and selection mechanisms to identify causal models from data and to clarify the nature of identifying assumptions. It is the dependence of unmeasured determinants of treatment choices with unmeasured determinants of potential outcomes that gives rise to the selection bias in empirically constructing counterfactual and treatment effect, even after conditioning on the observables.

The treatment effect literature avoids many of the problems confronted in the econometrics literature that builds explicit models of counterfactuals and assignment mechanisms. This literature makes fewer statistical assumptions in terms of independence, functional form, exclusion restriction and

distributional assumption. At the same time, it produces parameters that are more limited in application. Without further assumptions, these parameters do not lend themselves to extrapolation out of sample or to accurate forecasts impacts of other policies besides the ones being empirically investigated. At the same time, this literature is often unclear in stating what economic questions are estimated parameters answers. Simplicity in estimation is often accompanied by obscurity in interpretation. This approach does not use information about basic behavioral parameters obtained from other studies. When the components of treatments vary across studies, knowledge does not accumulate across treatment effect studies.

The econometric models are criticized for the interpretability of the economic frameworks and the parameters derived from them. At the same time there are questions about the strong functional form, the exogeneity, the support and exclusion assumptions used in classical versions of this literature, and the lack of robustness of empirical results. The arbitrariness in the choice of parametric models motivates recent work in nonparametric and semi-parametric econometrics.

The treatment effect approach does not model the factors determining outcomes, it works solely with outcomes without inputs, differently from the econometric approach that uses data to generate counterfactual. In such way they focus treatment effects for policies actually experienced and provide no framework for extrapolation of findings to new environments (problem P2) or for forecasting new policies (problem P3). Forecasting the effects of new policies is a central task of science, in fact econometricians model $Y(s, \omega)$ in terms of characteristics of treatment, and of the treated. In this way they facilitate comparisons of counterfactuals and derive causal effects across studies where the composition of programs and treatment group members may vary. It also facilitates the construction of counterfactuals on new populations and for new policies.

Incorporating choice into the analysis of treatment effects is an essential and distinctive ingredient of the econometric approach to the evaluation of social programs. The traditional treatment-control analysis in statistics equates mechanisms a and τ. An assignment in that literature is an assignment to treatment, not an assignment of incentives or of eligibility for treatment with the agent making treatment choices. In this notation, the traditional approach has only one assignment mechanism and treats noncompliance with it as a problem rather than as a source of information on agent preferences, which is a central feature of the econometric approach (Heckman and Smith, 1998). Thus, under full compliance $a : \Omega \to \mathcal{S}$, and $a = \tau$, where $\mathcal{B} = \mathcal{S}$.

The statistical approach does not model the treatment assignment rule or its relationship to potential outcomes. The econometric approach makes the treatment assignment equation the centerpiece of its focus and considers both objective and subjective valuations as well as ex ante ($\mathbb{E}(Y_1|\mathcal{I})$,

$\mathbb{E}(Y0|\mathcal{I})$, $\mathbb{E}(C|\mathcal{I})$) and ex post outcomes (Y_1, Y_0, C). The factors that lead an agent to participate in treatment may be dependent on the factors affecting outcomes. Modeling this dependence is a major source of information used in the econometric approach to construct counterfactuals from real data.

Economists distinguish objective and subjective outcomes, hence the statistical literature focuses exclusively on ex post objective outcomes. Indeed, statisticians reason in terms of assignment mechanism, while economists recognize the agent preferences often governed by choice. Modeling this choice process is a distinctive feature of the econometric approach. In econometrics comparison across outcome can be made also in terms of personal utilities or in terms of planner preferences. Utility function in fact produce subjective valuations of outcomes by the agents being treated. Agent making decisions about treatment may be only partially informed about realized payoffs at the time they make decision, then it is interesting to model the distinction between anticipated and realized outcomes.

Since statisticians do not develop choice equations or subjective evaluations, they do not consider the general invariance conditions for both objective and subjective evaluations. Indeed Rubin (1986) invokes versions of traditional econometric invariance assumptions and calls it *SUTVA* : Stable Unit Treatment Value Assumption. This assumption can be written in the econometric notation as follows:

A-1 $Y(s, \omega, p, \tau) = Y(s, \omega, \tau)$ that means no social interactions or general equilibrium effects;

A-2 $Y(s, \omega, \tau) = Y(s, \omega)$ that means the outcome is the same no matter what the choice of assignment mechanism.

The second condition rules out the phenomenon called *randomization bias* by Heckman *et al.* (1999) where agent's behavior is affected by the act of participating in an experiment. Furthermore they discuss the evidence against both assumptions. These assumptions represent strong limitations and in recent work Heckman and Vytlacil (2005) relax them.

It is possible to summarize the main characteristics of the two approaches below.

The Rubin model assumes:

R-1 $Y(s, \omega)_{s \in S}$, a set of counterfactuals defined for ex-post outcomes. It does not analyze agent valuations of outcomes nor does it explicitly specify treatment selection rules, except for contrasting randomization with non randomization;

R-2 (A-1) Invariance of counterfactuals for objective outcomes to the mechanism of assignment within a policy regime;

	Newman-Rubin framework	**Structural framework**
Counterfactual for objective outcomes	Yes	Yes
Agent valuations of subjective outcomes	No (choice mechanism implicit)	Yes
Models for the causes of potential outcomes	No	Yes
Ex ante versus ex post counterfactuals	No	Yes
Treatment assignment rules that recognize the voluntary nature of participation	No	Yes
Social interaction, general equilibrium effects and contagion	No (assumed away)	Yes (modeled)
Internal validity (P1)	Yes	Yes
External validity (P2)	No	Yes
Forecasting effects of new policy (P3)	No	Yes
Distributional treatment effects	No	Yes
Analyzing relationship between outcomes and choice equations	No (implicit)	Yes (explicit)
Treatment of interdependence	Recursive	Recursive or Simultaneous system

Table 2.2: Comparison of the aspects of policy evaluation covered by the Neyman-Rubin approach and the structural approach

R-3 (A-2) No social interactions or general equilibrium effects for objective outcomes;

R-4 P1 is the only problem of interest;

R-5 Mean causal effects are the main object of interest;

R-6 There is no simultaneity in causal effects, i.e., outcomes cannot cause one another.

The econometric approach considers a wider array of policy problems than the statistical treatment effect approach. Its signature features are:

- Development of an explicit framework for outcomes $Y(s,\omega)_{s\in S}$, measurements and the choice of outcomes where the role of unobservables in creating selection problems and justifying estimators is explicitly developed.
- The analysis of subjective evaluations of outcomes $R(s,\omega)_{s\in S}$, and the use of choice and compliance data to infer them (way to introduce agent decision making).
- The analysis of ex ante and ex post realizations and evaluations of treatments. This analysis enables analysts to model and identify regret and anticipation by agents (way to introduce agent decision making).

- Development of models for identifying and evaluating entire distributions of treatment effects (ex ante and ex post) rather than just the traditional mean parameters. These distributions enable analysts to determine other parameter like the proportion of people who benefit from treatment, a causal parameter not considered in the statistical literature on treatment effects.
- Models for simultaneous causality.
- Definitions of parameters made without appeals to hypothetical experimental manipulations.
- Clarification of the need for invariance of parameters with respect to different classes of manipulations to answer different classes of questions (P1-P3).

2.3.1 Reconciling the two literature

Structural model make the preferences and constraints explicit which given individual decisions, that rule interaction among agents and the sources of variability across agents. These feature facilitate finding answers to more policy questions, absent in the program evaluation literature. In the statistical literature there is the absence of explicit model. Fewer assumptions in term of exogeneity, functional form, exclusion and distributional assumptions than the standard structural estimation literature in econometrics are attractive features of this approach. The greater simplicity of estimation favours replicability, transparency and sensitivity analysis. Despite the recent advances in the structural literature, fully-specified structural models are often still hard to compute.

Heckman to reconcile these two literatures goes back to the Marschak's (1953) paper. Marschak noted that for many specific questions of policy analysis, it is unnecessary to identify full structural models. In some situations the parameters required to forecast particular policy modifications are represented by a combination of subsets of the structural parameters, which are much easier to identify. They require fewer and weaker assumptions. For example, policy that only affects X may be forecast using reduced forms, not knowing the full structure. Otherwise structural means parameters invariant to classes of policy modifications can be used. Thus to forecast other policies requires a partial knowledge of the system. Heckman called this principle *Marschak's Maxim* in honour of this insight and interpreted the modern statistical literature as implicitly implementing this principle. In such case the policy analyzed is the treatments and the goal of policy analysis is restricted to evaluating policy in place (P1).

Population mean treatment parameters are often identified under weaker conditions than the traditionally econometric analysis. To identify the aver-

age treatment effect is only required $E[Y(s,\omega)|X = x] - E[Y(s',\omega)|X = x]$. One does not have to know the full functional form of the generating g_s functions nor does X have to be exogenous. The treatment effects may, or may not, be causal parameters depending on what else is assumed about the model. Heckman (Heckman, 2005, 2010) considers identification conditions that underlie matching and instrumental variable methods. Moreover he discusses sources of unobservables, implicit assumptions about how unobservables are eliminated as source of selection problems, and the assumed relationship between outcomes and choice equation. He uses economics to interpret the parameter estimated and to make them useful for evaluating a wider range of policies.

Chapter 3

Dynamic model

This chapter surveys the main dynamic models which are referred to longitudinal and panel data. They have been recently proposed in the literature to face the fact that a treatment or a policy may be evaluated dynamically on time when such type of data are available. They allow to control unobserved heterogeneity and to model changes in time. This chapter proposes a survey of the main dynamic models. The importance of panel data and the advantages of dynamic models are explained. The most used models in literature are described under the main assumptions of dynamic unconfoundedness and of unobservables as for the static models described in the previous chapters. The dynamic model with the assumption of dynamic selection on observable represents the dynamic extension of the Newman-Rubin potential outcome model summarized in the first chapter. Among the models with the assumptions of selection on unobservable a model for categorical outcome variable and continuous outcome variables is presented. The former presents the dynamic binary model and the latter the duration and event-history models.

3.1 Advantages of dynamic model

In the last decade panel data have become widely available. For a given individual multiple observations are available, also in different points in time. Given this information a researcher is able to identify an otherwise unidentified model and simplify the computation and inference. Such data have several major advantages over conventional cross-sectional or time-series data sets. For example, they solve the problem arising in the presence of unobserved variables that are correlated with explanatory observable variables. By utilizing information on both the intertemporal dynamics and the individuality of the entities being investigated, one is better able to control in a natural way for the effects of such variables. In policy context, it is expected that a policy can affect not only current outcomes, but also outcomes

at other points in time. Also the same policy implemented at different time periods may have different consequences. Having panel data on past program participation, intermediate outcomes and covariates can be useful to the assignment of programs and to the control of dynamic selection process. The static framework provides a considerable simplification as compared to a dynamic setting, but it does not account for attrition from the proposed treatment while it is in operation. In some evaluation studies it may be possible to phrase the problem and organize data so that it fits the static setup or by using ad hoc modification of the static causal framework. More often the proper economic interpretation of parameters and identifying assumptions is hard if a dynamic problem is framed as a static problem. Standard statistical approach may fail to estimate or test useful parameters. In epidemiology and biostatistics similar problems are faced using counterfactual outcomes explicitly in the dynamic setting in the analysis of the casual effects of complex dynamic medical treatments on health outcomes. In such context the main approach used is the dynamic versions of the potential outcome approach, suggested by Robins (1986) and developed in Robins (1989, 1997), Gill and Robins (2001), and Lechner and Miquel (2010). These papers are based on the so-called selection on observables assumption. Lechner (2004) applies the framework to the labor market programs and Abbring (2008) extends the dynamic model including agent choice and subjective evaluations (For a survey of these models see Abbring and Heckman (2008)). Explanation of this model can be found in the following paragraphs. Often, in econometric program evaluations selection on observables assumption is unlikely to hold, because usually observational data are characterized by a lot of heterogeneity among agents, not fully captured by the observed variables. In a dynamic context, such unmeasured heterogeneity leads to violations of that assumption. This is true even if the unmeasured variables do not affect outcomes directly, because if agents are rational, forward-looking and observe at least some of the unmeasured variables that the econometrician does not, they will typically respond to these variables through their choice of treatment and investment behavior. For the same reason, identification based on instrumental variables is relatively hard to justify in dynamic models (Hansen and Sargent, 1980; Rosenzweig and Wolpin, 2000; Abbring and Van den Berg, 2005). If the candidate instruments only vary across persons but not over time for the same person, then they are not likely to be valid instruments because they affect expectations and future choices and may affect current potential outcomes. Instead of using instrumental variables that vary only across persons, instruments based on unanticipated person-specific shocks that affect treatment choices but not outcomes at each point in time are required. The econometric literature is rooted on state dependence and heterogeneity. Any history dependence of transition rates can be explained both as true state dependence and as the result of unobserved heterogeneity that simultaneously affects the history and current transitions.

This is a dynamic manifestation of the problem of drawing causal inference from observational data.

3.1.1 State dependence, heterogeneity and initial condition

When the conditional probability of an individual staying in a state is a function of past experience, two new issues arise: how to treat initial observation and how to distinguish true state dependence from spurious state dependence. In much applied work usually initial conditions of relevant history of the process are assumed to be truly exogenous (Hsiao, 2003). This means that initial conditions are fixed. Such hypothesis may be justifiable only if disturbances that generate the process are serially independent or if a new process is observed at the beginning of the sample. If the process has been in process prior to the time it is observed, or if the disturbances of the model are serially dependent, the initial conditions are not exogenous. For example, the initial state cannot be assumed fixed in the presence of individual specific random effects. In this case the marginal distribution of initial condition given the individual specific random effect has to be derived. This is not so easy since the initial state is a function of unobserved past values. For more details see Hsiao (2003). For what concerns state dependence, the problem is to distinguish between true and spurious state dependence. Both of them can explain why individuals who have experienced an event in the past are more likely to experience that event in the future than individuals who have not experienced the event. In fact there are two main opposite explanations for this empirical regularity:

1. individuals are influenced by the experience of the event, preferences or constraints relevant to future choice change. In such a case identical individual who has not experienced the event will behave differently in the future than an individual who has experienced the event;

2. individuals differ in certain unmeasured variables that influence their probability of experiencing the event. If these variables are correlated over time and are not controlled, previous experience may appear to be a determinant of future experience. This happens because the previous experience is a proxy for temporally persistent unobservables that determine the choice.

Heckman (Heckman, 1981) called the former *true state dependence* and the latter case *spurious state dependence*. This problem is very similar to the econometric problem of estimating a lag model in the presence of serial correlation. In this case to account for heterogeneity the error term ϵ_{it} in the model is decomposed as follows:

$$\epsilon_{it} = \alpha_i + u_{it} \quad i = 1, \ldots, N \quad t = 1, \ldots, T,$$

where u_{it} is independently distributed over i with arbitrary serial correlation, and α_i is a subjective-specific variable and can be treated as a fixed constant or as random. A model specification in which are considered past experiences and the error term specified as above identifies three sources of persistence:

- serial correlation in the error term u_{it};
- unobserved heterogeneity α_i;
- true state dependence given by lagged variables $y_{i,t-1}$.

For further details see Heckman (1981), Hsiao (2003) and Bartolucci (2010) Lagged variables capture the state dependence while the unobserved individual effects (α_i) controlled the serial correlation of errors. Conditional on the individual effect α_i, the error term u_{it} should be serially uncorrelated. On the contrary, if the error term u_{it} remains serially correlated then past y_{it} contains information on u_{it} and the problem becomes more complicated. In most papers the heterogeneity is assumed to follow a continuous distribution. The most commonly assumed distributions used are the Normal, Lognormal and Gamma distribution [See Lancaster (1990)]. An important limitation of this approach is that its estimation relies on an a priori assumption about the shape of the distribution, which might lead to biased estimates in cases where the underlying distribution has a substantially different shape from the assumed distribution (Heckman and Singer, 1984). An alternative approach is to assume that the distribution is non-parametric, where the distribution of parameters is represented by a finite number of mass points. It is a discrete representation of heterogeneity, as the support of the distribution is discrete. This mass point approach can be seen as closely related to the Latent Class model (LCM), also called the Finite Mixture model (McLachlan and Peel, 2000), as it implies that the parameters influencing different individuals are associated with membership in a distinct class or group.

3.2 Dynamic extension for the Newman-Rubin potential outcome model

3.2.1 The notation

The dynamic policy evaluation problem can be formalized in a fashion similar to the way in which the static problem is formalized.

The possible treatment assignment times are $1, \ldots, \overline{T}$. The set $\mathcal{S}$ of treatments considered is not restricted and it is allowed to the same treatment to be assigned on multiple occasions. For expositional convenience, $\mathcal{S}$ is supposed to be a finite discrete set. In general, the set of available treatments

at each time t may depend on time t and on the history of treatments, outcomes, and covariates. A dynamic policy $p = (a, \tau) \in \mathcal{A} \times \mathcal{T} \equiv \mathcal{P}$ is defined as a dynamic constraint assignment rule $a = \{a_t\}_{t=1}^{\overline{T}}$ with a dynamic treatment choice rule $\tau = \{\tau_t\}_{t=1}^{\overline{T}}$. I indicate the sequence of events until time t with the apix t (for example $a_p^t = (a_p(1), \ldots, a_p(t))$), while the whole sequence of event until T without apex (for example $a_p = (a_p(1), \ldots, a_p(\overline{T}))$). Agent, policy's planner and econometrician (or analyst), at time t under policy p, have a different information set, respectively $\mathcal{I}_A(t,p), \mathcal{I}_P(t,p), \mathcal{I}_E(t,p)$:

- each agent ω chooses treatment $\tau_t(\omega, a)$ given determinants of constraints, future outcomes and the own preference components.
- planner assigns constraints to each agent using information based on covariates and random variables under the planners control, as well as past choices and realized outcomes.
- econometrician observes external covariates, not affected by the policy p, Z of the assignment mechanism and X of the potential outcome. The unobserved external covariates by the econometrician are indicated with U_t, $t = 1, \ldots, \overline{T}$.

At each time t, each actor acts using the information about policy p at time t present in his information set $\mathcal{I}(t,p)$. Furthermore it is assumed that the information is not forgotten, so each agent improves his information over time and $\mathcal{I}(t,p) \subseteq \mathcal{I}(t+1,p)$ for all t. It is also assumed that agents know more than the planner at each time t, so that $\mathcal{I}_P(t,p) \subseteq \mathcal{I}_A(t,p)$. Objective outcomes associated with policies p are expressed as a vector of time-specific outcomes $Y_p = (Y_p(1), \ldots, Y_p(\overline{T}))$.

Extending the notation for the static case, the assignment of agents to treatment it is denoted by $s_p(\omega, t) = \tau_t(\omega, a)$. The shorthand s_p^t is used for the vector $(s_p(1), \ldots, s_p(t))$ of treatments assigned up to and including time t under policy p, and s_p is used to indicate $s_p^{\overline{T}}$. Treatment assignment $s_p(t)$ is then a random variable that only depends on the agents information. To make this dependence explicit, suppose that external covariates Z and unobserved external covariates V_1 that affect treatment assignment are revealed to the agents at time 1. It follows that the information set of the agent $\mathcal{I}_A(1,p)$ at the different periods is:

- for $t = 1 \quad \mathcal{I}_A(1,p) = \sigma(Z, V_1)$
- for $t \geq 2 \quad \mathcal{I}_A(t,p) = \sigma(Y_p^{t-1} Z, V^t)$.

The initial information set of the agents $\mathcal{I}_A(1,p)$ includes only external variables, so it does not depend on policy p. The ex post potential outcomes corresponding to each treatment sequence $s = (s(1), \ldots, s(\overline{T}))$ are $Y(t,s) = y_t(s, X, U_t)$ with $t = 1, \ldots, \overline{T}$.

3.2.2 Assumptions

To simplify the notation, in the following text the dependence of outcomes on observed covariates X is assumed implicit and all conditioning on X is suppressed. A first assumption to the identification of effect in a dynamic context is that there is no causal dependence of outcomes on future treatment. That means:

$$\text{NA for all} \quad t \geq 1, Y(t,s) = Y(t,s') \quad \text{for all} \quad s, s' \quad \text{such that} \quad st = (s')^t,$$

where $s_t = (s(1), \ldots, s(t))$ and $(s')^t = (s'(1), \ldots, s'(t))$. Abbring and Van den Berg (2003b) and Abbring and Van den Berg (2003a) define this as a *no-anticipation condition* (NA). It requires that outcomes at time t (and before) be the same across policies that allocate the same treatment up to and including t, even if they allocate different treatments after t. In the structural econometric models this condition is trivially satisfied if all state variables relevant to outcomes at time t are included as inputs in the outcome equations $Y(t,s) = y_t(s, U_t), t = 1, \ldots, \overline{T}$.

Condition of NA implies that actual outcomes up to time $t-1$ are equal between two different policies p and p' ($Y_p^{t-1} = Y_{p'}^{t-1}$), if the treatment history coincide up to time $t-1$ ($s_p^{t-1} = s_{p'}^{t-1}$). The treatment choice is then determined by the distributional properties of the constraint assignment rule a, as the past and current constraints that were actually assigned to him and by agent ω's predictions of future constraints and outcomes. A forward-looking agent will use observations of his covariates $Z(\omega)$ and $V_t(\omega)$ and past outcomes $Y_p^{t-1}(\omega)$ to infer his type ω and subsequently predict future external determinants $(U_t(\omega), \ldots, U_{\overline{T}}(\omega))$ of his outcomes and $(V_t(\omega), \ldots, V_{\overline{T}}(\omega))$ of his constraints and treatments. In turn, this information updating allows agent to predict his future potential outcomes and, for a given policy, his future constraints, treatments and realized outcomes. Justifying the (NA) assumption requires specification of agent information about future treatment and agent behavior in response to that information. Even though the time t agent information about ω is the same under both policies, agents may have different predictions of future constraints, treatments and outcomes because the policies may differ in the future and agents know this. (NA) requires potential outcomes to be determined externally, and not to be affected by agent actions in response to different predictions of future constraints, treatments and outcomes. For this reason it is necessary to consider also the effect of the information available to agents about the program and policy. The analyst needs to control for the effect of agents' information modelling it.

To identify the dynamic treatment effects, Gill and Robins (2001) invoke a dynamic version of the matching assumption (conditional independence) which relies on *sequential randomization* (SR) for all treatment sequences s and all t is:

$$\text{SR } S(t) \perp\!\!\!\perp (Y(t,s),\ldots,Y(\overline{T},s))|(Y_{p_0}^{t-1}, S^{t-1}=s^{t-1}, Z),$$

where the conditioning set $(Y_{p_0}^0, S^0 = s^0, Z)$ for $t = 1$ is Z. Equivalently, $S(t) \perp\!\!\!\perp (U_t,\ldots,U_{\overline{T}})|(Y_{p_0}^{t-1}, S^{t-1}, Z)$ for all t without further restricting the data. Sequential randomization allows the $Y_{p_0}(t)$ to be *dynamic confounders* variables that are affected by past treatment and that affect future treatment assignment. The sequence of conditioning information sets appearing in the sequential randomization assumption represent the econometrician information set:

- for $t = 1$ $\mathcal{I}_E(1) = \sigma(Z)$
- for $t \geq 2$ $\mathcal{I}_E(t) = \sigma(Y_{p_0}^{t-1}, S^{t-1}, Z)$.

Furthermore it is supposed that $\mathcal{I}_E(t) \subseteq \mathcal{I}_A(t, p_0)$ for each t. If treatment assignment is based on strictly more information than $\mathcal{I}_E$, so that agents know strictly more than the econometrician and act on their superior information, SR is likely to fail if that extra information also affects outcomes. No anticipation and sequential randomization represent the identifying assumptions. They are not testable because they do not restrict the factual data (Gill and Robins, 2001).

3.2.3 Identification of causal effect: the *g*-computation formula

The sequential randomization SR together with the no-anticipation condition (NA) represent the natural dynamic extension of the Neyman-Roy-Rubin model for a static randomized experiment. Under such assumptions and through the *g-computation formula* it is possible to sequentially identify the causal effects of treatment from the distribution of the data (Y_{p_0}, S, Z) and construct the distribution of the potential outcomes $Y(s)$ for any treatment sequence s in the support of S. To explain the Robins (1997)'s *g*-computation formula I will consider the case in which all variables are discrete. No-anticipation condition (NA) ensures that potential outcomes for a treatment sequence s equal outcomes under policy p_0 (actual policy) up to time $t-1$ for agents with treatment history s_{t-1} up to time $t-1$. Formally, $Y^{t-1}(s) = Y_{p_0}^{t-1}$ on the set $S_{t-1} = s_{t-1}$. Using this, sequential randomization assumption SR can be rephrased in terms of potential outcomes: for all s and t,

$$S(t) \perp\!\!\!\perp (Y(t,s),\ldots,Y(\overline{T},s))|(Y_s^{t-1}, S^{t-1}=s^{t-1}, Z).$$

In turn, this implies that, for all s and t,

$$P(Y(t,s)=y(t)|Y_s^{t-1}=y^{t-1}, S^t=s^t, Z) = P(Y(t,s)=y(t)|Y_s^{t-1}=y^{t-1}, Z)$$

where $y_{t-1} = (y(1), \ldots, y(t-1))$ and $y = y^{\overline{T}}$. From Bayes rule and the above equation, it follows that:

$$P[Y(s) = y|Z] = P[Y(1,s) = y(1)|Z] \prod_{t=2}^{\overline{T}} P[Y(t,s) = y(t)|Y^{t-1}(s) = y^{t-1}, Z] =$$

$$= P[Y(1,s) = y(1)|S(1), Z] \prod_{t=2}^{\overline{T}} P[Y(t,s) = y(t)|Y^{t-1}(s) = y^{t-1}, S^t = s^t, Z]$$

Invoking (NA), in particular $Y(t,s) = Y_{p_0}(t)$ and $Y^{t-1}(s) = Y_{p_0}^{t-1}$ on $S^t = s^t$, produces

$$P[Y(s) = y|Z) =$$

$$= P(Y_{p_0}(1) = y(1)|S(1), Z] \prod_{t=2}^{\overline{T}} P[Y_{p_0}(t) = y(t)|Y_{p_0}^{t-1} = y^{t-1}, S^t = s^t, Z]$$

This is the version of Robins's (1997) g-computation formula. From data it is possible to sequentially identify each component on the left hand side of the first expression, and hence identify the counterfactual distributions. Matching on pretreatment covariates is a special case of the g-computation approach in a static model. See Abbring (2008) for further details. An alternative approach is to explicitly model and identify the evolution of the unobservables.

3.3 The dynamic binary model and its extension

When the dependent variable Y is a discrete variable that represents a category, from a set of mutually exclusive categories then logit, nested logit, and probit models are used to model the relationship with one or more independent variables X (see among others Agresti (2002)). Among the statistical and econometric model for binary longitudinal data, the *dynamic logit model* (Hsiao, 2003) is of particular interest and finds applications in many fields as in the study of the labour market.

Let y_{it} denote the binary response variable for subject i at occasion t, with $i = 1, \ldots, n$ and $t = 1, \ldots, T$ and let x_{it} be a corresponding vector of strictly exogenous covariates. The dynamic logit model (Agresti, 1990) assumes that:

$$\log \frac{p(y_{it} = 1|\alpha_i, \mathbf{x}_{it}, y_{i,t-1})}{p(y_{it} = 0|\alpha_i, \mathbf{x}_{it}, y_{i,t-1})} = \alpha_i + \mathbf{x}_{it}'\beta + y_{i,t-1}\gamma$$

where α_i is a subject-specific parameter which captures the unobserved covariates, β is a vector of regression coefficients for the covariates and γ is the parameter of state dependence. This model is justified in econometric

literature on the basis of the structural model proposed by Heckman (1981) in term of a continuous latent random variable crossing a threshold. In such case the discrete outcome y can be viewed as the unobserved counterpart of a continuous latent (unobservable) variable y^*_{it}:

$$y^*_{it} = \mathbf{x'}_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \epsilon_{it}, \qquad t = 1, \ldots, T$$

The error term ϵ_{it} is assumed to be independent of $\mathbf{x'}_{it}$ and over i and to follow a logistic distribution. The observed variable can be written as:

$$y_{it} = \begin{cases} 1 & \text{if} \quad y^*_{it} > 0 \\ 0 & \text{if} \quad y^*_{it} \leq 0 \end{cases}$$

Given that y is a categorical outcome variable with $j = 1, \ldots, J$ numbers of levels (or categories), the marginal distribution of this variable may be described by a set of $J - 1$ logits chosen from among four different types according to characteristics of such categories: (Colombi and Forcina, 2001):

- local (l) : $\log[P(Y = j + 1)] - \log[P(Y = j)]$
- global (g) : $\log[P(Y > j)] - \log[P(Y \leq j)]$
- continuation (c) : $\log[P(Y > j)] - \log[P(Y = j)]$
- reverse continuation (r) : $\log[P(Y = j + 1)] - \log[P(Y \leq j)]$.

Logit of type g and c are the most appropriate for ordered categorical responses. The global one can be used to construct logits of cumulative probabilities. Given $\pi_j = P(Y = j)$ then:

$$\text{logit}[P(Y \leq j)] = \log \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \log \frac{P(Y \leq j)}{P(Y > j)} = \log \frac{pi_1 + \ldots + pi_j}{pi_{j+1} + \ldots + pi_J}.$$

When there are more then one variable of interest and they are categorical variables with more then two level extensions of this model has to be used. Extension and application of this model can be found in Bartolucci and Farcomeni (2009) and Bartolucci and Pennoni (2010).

Let r denote the number of categorical response variables observed at each occasion and denote by y_{hit} the h-th response variable for subject i at occasion t, with $h = 1, \ldots, r$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. This variable has l_h categories indexed from 0 to $l_h - 1$. Let $\mathbf{y_{it}}$ denote the vector with elements y_{hit} and let $p(\mathbf{y_{it}}|\boldsymbol{\alpha_{it}}, \mathbf{x_{it}}, \mathbf{y_{i,t-1}})$ the conditional probabilities for all possible configuration of y_{it} arranged in order, where $\mathbf{x_{it}}$ denote the covariates, $\mathbf{y_{i,t-1}}$ the lagged response and $\boldsymbol{\alpha}_{it}$ the time-varying random effects. A model assumption that can be made is that y_{it} is conditional independent of $y_{i0}, \ldots, y_{i,t-2}$, given $x_{it}, y_{i,t-1}$ and α_{it}, $t = 2, \ldots, T$. As an example consider the case of three variables ($r = 3$) with 2, 3 and 3 levels ($l_1 = 1, l_2 = 2, l_3 = 3$),

respectively. Furthermore, the logit can be treated with logits of type local, global and continuation, respectively. The logit can be then parametrized as follows:

$$\log \frac{p(y_{hit} = 1|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{i,t-1})}{p(y_{hit} = 0|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{\mathbf{i,t-1}})} = \alpha_{1it} + \mathbf{x}_{it}^{'}\boldsymbol{\beta}_1 + \mathbf{y}_{i,t-1}^{'}\boldsymbol{\gamma}_1$$

$$\log \frac{p(y_{hit} \geq z|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{i,t-1})}{p(y_{hit} < z|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{i,t-1})} = \alpha_{z+1,it} + \mathbf{x}_{it}^{'}\boldsymbol{\beta}_2 + \mathbf{y}_{i,t-1}^{'}\boldsymbol{\gamma}_2 \quad z = 1, 2$$

$$\log \frac{p(y_{hit} \geq z|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{i,t-1})}{p(y_{hit} = z-1|\boldsymbol{\alpha}_{it}, \mathbf{x}_{it}, \mathbf{y}_{i,t-1})} = \alpha_{z+3,it} + \mathbf{x}_{it}^{'}\boldsymbol{\beta}_3 + \mathbf{y}_{i,t-1}^{'}\boldsymbol{\gamma}_3 \quad z = 1, 2$$

Note that on one hand the regression coefficients for the covariates and those for the lagged response variables are the same for both logits ($z = 1, 2$) and on the other hand the intercept α_{hit} are specific to each variable category. This is a standard practice in model for ordinal variable (McCullagh, 1980). The latent process or heterogeneity $\alpha_{i1}, \ldots, \alpha_{iT}$ can be modeled in different ways as explained in the previous paragraph.

3.4 Treatment effect in duration model

Let us consider two continuously-distributed random durations: S, the duration of the time to treatment and Y, the outcome duration. Let $Y(s)$ and $S(y)$ respectively be the potential outcome duration when the treatment time is externally set to s and the potential treatment time resulting from setting the outcome duration to y. It is assumed that ex ante heterogeneity across agents is fully captured by observed covariates X and unobserved covariates V, assumed to be external and temporally invariant. Treatment causally affects the outcome duration through its hazard rate. The hazard rate of $Y(s)$ at time t for an agent with characteristics (X, V) is denoted by $\theta_Y(t|s, X, V)$. Similarly, outcomes affect the treatment times through their hazard $\theta_S(t|y, X, V)$. Without loss of generality, it is possible to partition V into (V_S, V_Y) and assume that:

$$\begin{aligned} \theta_Y(t|s, X, V) &= \theta_Y(t|s, X, V_Y) \\ \theta_S(t|y, X, V) &= \theta_S(t|y, X, V_S) \end{aligned} \tag{3.1}$$

Intuitively, V_S and V_Y are the unobservables affecting, respectively, treatment and outcome, and the joint distribution of (V_S, V_Y) is unrestricted. In particular, V_S and V_Y may have elements in common. The corresponding

integrated hazard rates are:

$$\Theta_Y(t|s,X,V_Y) = \int_0^t \theta_Y(u|s,X,V_Y)du$$

$$\Theta_S(t|y,X,V_S) = \int_0^t \theta_S(u|y,X,V_S)du \tag{3.2}$$

For expositional convenience, it is assumed that these integrated hazards are strictly increasing in t and that they diverge to ∞ as $t \to \infty$, so that the duration distributions are non-defective. Then, $\Theta_S(t|y,X,V_S)$ and $\Theta_Y(t|s,X,V_Y)$ are unit exponential for all $y,s \in \mathbb{R}_+$. This implies the following model of potential outcomes and treatments

$$Y(s) = y(s,X,V_Y,\epsilon_Y) \quad S(y) = s(y,X,V_S,\epsilon_S), \tag{3.3}$$

for some unit exponential random variables ϵ_Y and ϵ_S that are independent of (X,V), $y = \Theta_Y^{-1}$ and $s = \Theta_S^{-1}$. The exponential errors ϵ_Y and ϵ_S represent the randomness in the transition process after conditioning on covariates X and V and survival and are assumed to be independent $\epsilon_Y \perp\!\!\!\perp \epsilon_S$. In such way $Y(s)$ and $S(y)$ are only dependent through the observed and unobserved covariates (X,V).

This *conditional-independence assumption* is weaker than the conditional-independence assumption used in matching, because it allows for conditioning on the invariant unobservables V. It is also assumed a version of the *no-anticipation condition*: for all $t \in \mathbb{R}_+$,

$$\theta_Y(t|s,X,V_Y) = \theta_Y(t|s',X,V_Y)$$

$$\theta_S(t|y,X,V_S) = \theta_S(t|y',X,V_S) \tag{3.4}$$

for all $s,s',y,y' \in [t,\infty)$.

This excludes effects of anticipation of the treatment on the outcome. Similarly, there can be no anticipation effects of future outcomes on the treatment time hazard. The no-anticipation assumption ensures that this system has a unique solution (Y,S) by imposing a recursive structure on the underlying transition processes. Together with a distribution $G(\cdot|X)$ of $V|X$, this gives a non-parametric structural model of the distribution of $(Y,S)|X$ that embodies general simultaneous causal dependence of Y and S, dependence of (Y,S) on observed covariates X, and general dependence of the unobserved errors V_Y and V_S.

There are two reasons for imposing further restrictions on this model. Firstly, it is not identified from data on (Y,S,X). For each version of the model with selection on unobservables and anticipation effects, there is an observationally-equivalent model version that satisfies no-anticipation and

conditional randomization. Secondly, even if its ensured nonparametric identification by assuming no-anticipation and conditional randomization, it is possible to learn, at best, only about the average effects:

$$\theta_Y(t|s,X) = \mathbb{E}[\theta_Y(t|s,X,V_Y)|X,Y(s) \geq t]$$

$$\theta_S(t|y,X) = \mathbb{E}[\theta_S(t|y,X,V_S)|X,Y(s) \geq t] \tag{3.5}$$

Thus, it is possible to identify the distributions of $Y(s)|X$ and $S(y)|X$. These distributions do not control for unobserved covariates V, so they can lead up to dynamic confounding selection effects (Abbring and Van den Berg, 2005).

Abbring and Van den Berg study the model identifiability without exclusion restrictions considering an extension of the multivariate *Mixed Proportional Hazard* (MPH) model (Lancaster, 1990) in which the hazard rates of $Y(s)|(X,V)$ and $S(y)|(X,V)$ are given by

$$\theta_Y(t|s,X,V) = \begin{cases} \lambda_Y(t)\phi_Y(X)V_Y & \text{if } t \leq s \\ \lambda_Y(t)\phi_Y(X)\delta_Y(t,s,X)V_Y & \text{if } t > s \end{cases}$$

and

$$\theta_S(t|y,X,V) = \begin{cases} \lambda_S(t)\phi_S(X)V_S & \text{if } t \leq y \\ \lambda_S(t)\phi_S(X)\delta_S(t,y,X)V_S & \text{if } t > y \end{cases}$$

respectively, and $V = (V_S, V_Y)$ is distributed independently of X. The baseline hazards $\lambda_Y : \mathbb{R}_+ \to (0,\infty)$ and $\lambda_S : \mathbb{R}_+ \to (0,\infty)$ capture duration dependence of the individual transition rates. The integrated hazards are $\Lambda_Y(t) = \int_0^t \lambda_Y(\tau)d\tau < \infty$ and $\Lambda_S(t) = \int_0^t \lambda_S(\tau)d\tau < \infty$ for all $t \in \mathbb{R}_+$. The regressor functions $\phi_Y : \mathcal{X} \to (0,\infty)$ and $\phi_S : \mathcal{X} \to (0,\infty)$ are assumed to be continuous, with $\mathcal{X} \subset \mathbb{R}^q$ being the support of X. In empirical work, these functions are frequently specified as $\phi_Y(x) = exp(x'\beta_Y)$ and $\phi_S(x) = exp(x'\beta_S)$ for some parameter vectors β_Y and β_S. The functions δ_Y and δ_S capture the causal effects. Note that $\delta_Y(t,s,X)$ only enters $\theta_Y(t|s,X,V)$ at durations $t > s$, so that the model satisfies no anticipation of treatment assumption (NA). Similarly, it satisfies no anticipation of outcomes and has a recursive causal structure as required by the no anticipation assumption. If $\delta_Y = 1$, treatment is ineffective; if δ_Y is larger than 1, it stochastically reduces the remaining outcome duration.

The MPH restriction on this model, however, is hard to justify from economic theory, indeed the MPH model only results under strong assumptions (Heckman and Singer, 1986; Van den Berg, 2001). This model allows δ_Y and δ_S to depend on elapsed duration t, past endogenous events, and the observed covariates X, but not on V. Abbring and Van den Berg (2003a)

also consider an alternative model that allows δ_Y and δ_S to depend on unobservables in a general way, but not on past endogenous events. They also show that these models are non parametrically identified from single spell data under the conditions for the identification of competing-risks models based on the multivariate MPH model. The models can be parameterized in a flexible way and estimated by maximum likelihood. Typical parameterizations involve linear-index structures for the regressor and causal effects, a discrete distribution G, and piecewise-constant baseline hazards for λ_S and λ_Y. An empirical application can be found in Abbring *et al.* (2005).

3.4.1 Treatment effect in more general event history models

It is possible to place the causal duration models in the more general setting of event-history models with state dependence and heterogeneity. A way to do this is to follow Abbring and Heckman's (2008) analysis of the mixed semi-Markov model, that is also analogous to the frameworks of Heckman and Singer (1986). The point of departure is a continuous-time stochastic process assuming values in a finite set $\mathcal{S}$ at each point in time. The realizations of this process are interpreted as agents event histories of transitions between states in the state space $\mathcal{S}$. Suppose that event histories start at real-valued random times T_0 in an S-valued random state S_0, and that subsequent transitions occur at random times $T_1, T_2, \ldots$such that $T_0 < T_1 < T_2 < \ldots$. Let S_l be the random destination state of the transition at T_l. Taking the sample paths of the event-history process to be right-continuous, S_l is the state occupied in the interval $[T_l, T_{l+1})$. Suppose that heterogeneity among agents is captured by vectors of time-constant observed covariates X and unobserved covariates V. In this case, state dependence in the event-history process for given individual characteristics X, V has a causal interpretation.

A way to give a structure to such state dependence it is to assume that the event-history process conditional on X, V is a *time-homogeneous semi-Markov process*. Conditional on X, V the length of a spell in a state and the destination state of the transition ending that spell depend only on the past states through the current state. In the potential outcome notation:

$$(\Delta T_l, S_l) \perp\!\!\!\perp (T_i, S_i), i = 0, \ldots, l-1 | S_{l-1}, X, V$$

where $\Delta T_l = T_l - T_{l-1}$ is the length of spell l. Also, the distribution of $(\Delta T_l, S_l)|S_{l-1}, X, V$ does not depend on l. Note that, conditional on X,V, $S_l, l \geq 0$ is a time-homogeneous Markov chain under these assumptions.

The event-history process conditional on observed covariates X only is a *mixed semi-Markov process*. If V affects the initial state S_0, or transitions from it, subpopulations of agents in different states at some time t typically have different distributions of the unobserved characteristics V. Therefore, a comparison of the subsequent transitions in two such subpopulations does

not only reflect state dependence, but also sorting of agents with different unobserved characteristics into the different states they occupy at time t. Abbring (2008); Abbring and Heckman (2007) discusses this model's application to program evaluation.

Chapter 4

Labour market and available data

This work attempts to investigate the dynamic impact of the first stable job coherent with the university education on the future job coherence. That is possible having at disposal administrative panel data on both Lombardy labour market and Milan Universities education. Before proceeding by modeling such a problem, a brief introduction of the nowadays labor market in Italy is presented. In the last decade global economic system has changed and with it also the concepts of job and stability. In this context it is of great relevance to redefine the concept of a good job. Considering the Subsample of graduates it is of great interest the duration and the coherence of the work experience with one's own studies. In this chapter the available administrative data are described and descriptive statistics are presented for the sample considered. Furthermore two subpopulations are taken into account: the subsample of people with at least a long term coherent job and the subsample of people with at least a stable job. Finally the main differences of this Subsamples compared with the whole population are highlighted.

4.1 Labour market

4.1.1 Italian labour market today

One of the focal points of the nowadays labour market and of the intervention of institutions can be found in the dialectic combination between two poles:

"On one hand the need for security and on the other the experience of change. The need for security, that is a perspective for a long term job, as working is planning the future, building something. The experience of the risk of change, because techniques change, needs to be answered; change and organizational contexts are constantly evolving "(Marco Martini, 1998).

Finding a balance between these two divergent trends is a vital need

for the harmonious development of the labor market. The global economic system has changed considerably in recent decades facing strong pressures like financing the economy, globalization, development of services and the ever more massive and rapid adoption of innovations derived from ICT. All this has undermined the *Fordist* productive system which dominated the economic organization since the mid-nineteenth century. The economic system is changing and with it business and workers needs.

On one hand, the unpredictability of the market and the increasing variability of demand put in a difficult position the original set of companies which try to react reorganizing their own activities. On the other hand, people lose the certainty of a life-long lasting workplace and a career that develops within a company with a known location and well defined growing path.

Given the rapid spread of temporary employment and the increased instability of the market, a new concept is taking root, the *work path*, which can take place in different sectors and positions and require very different skills and knowledge also quite different one from the other. Certainty, which was previously linked to the presence of stable, productive organization, lacks in this new context of the labour market. In a situation like this a strategic role is played by public, private and non-profit organizations, dealing with employment services, which more and more often interpose themselves in the process of matching job demands and job offers. Among these, temporary employment agencies, able to meet the needs of business flexibility. Work relationships become more flexible with the consequent growth of "non standard" work types. This term typically refers to all types of work that are not characterized by three components: full time engagement, presence of a single employer and a permanent contract. In addition to this, the workers, even those employed on a permanent contract, can no longer expect to remain employed for a long period in the same company. Technological and organizational change help, in fact, in significantly reducing the average length of employment relationships, even those standards, and higher rates of total turnover in the labour market (Bauer and Bender, 2004).

4.1.2 A new concept of career

What has characterized the labour market in recent years is a substantial increase of mobility at all levels. In the first place occupational mobility increases, in the meaning of the transition from one job to another or from one organization to another. Also professional mobility increases, as the transition from one professional position to another, and contractual mobility, as the transition through various employment forms. Work paths then develop through a consistent inter-sectoral mobility, because of which the workforce can be called to change the field of activity several times during one's career.

Work experience also changes within the same organization. New strategies for flexible management change position (employee, independent, co-worker, etc..), length of the work relationship (permanent, fixed-in project, etc..), commitment time (part-time, full time), workplace (tele-work, etc..), and remuneration as a fundamental part of an employment relationship, making it more variable and personal and business results driven.

That is why career in the post-Fordist society undergoes radical changes. In particular, we are witnessing what is called the transition from a linear career model to a boudaryless career. The linear career typically develops through professional advancement that occurs within one or two companies, following a linear path of growth stages. The success of the organization depends on the individual worker and is measured by the number of promotions and the wage increase. The career without boundaries is instead defined as "a sequence of job opportunities that develop beyond the boundaries of a single organizational context" (DeFillippi, 1996). Following this, today it is almost unthinkable for people to focus exclusively on finding the *job* that lasts a lifetime within the same organization and that guarantees a slow and structured professional development. It is more likely for individuals to engage in a *career path* that can take place in very different branches and positions, and that requires skills and knowledge rather far apart. Given this context it is essential for an employee to be still able to experience a career grow path over time developing both in terms of income and professionalism. To be successful, people need to acquire knowledge and skills usable in the market and no longer only within a single company.

4.1.3 The concept of stability

In the labor market literature (Booth *et al.*, 2002) stability is expressed by the type of job contract. Usually the main contract categories are permanent employment, that is employment of unspecified duration, and fixed-term or temporary employment, that ends automatically, without any prior termination procedure, on the date either when the contractual term expires or when the contractually specified work is completed. This classification was used, for example, by Bonnal *et al.* (1997) and Gagliarducci (2005). A more detailed categorization implies to distinguish between fixed-term and temporary contract. A fixed-term contract is one which terminates on a specified date or on the occurrence of an event which is certain to happen on a particular date. A temporary contract is normally used when no end date is known and its termination is dependent on an event such as return from sick leave or maternity leave, or completion of a job. Furthermore, usually, a temporary contract has a shorter duration than fixed-term contract, and it is got through the Work Temporary Agency (WTA). Along with a steady growth of the temporary work, introduced in Italy in 1997 (Law 196 / 197), over the past thirty years has also raised the interest in the role played by

this form of employment in the labor market.

National and international literature attempt to study the effect of temporary work on the *work process*. On one hand temporary job may be preferable to an unemployment state, as it represents a good opportunity to enter the labor market. On the other hand, accepting a temporary job could have a negative effect on the future perspectives of career stabilization and on levels of retention and motivation of employees. Temporary jobs can represent "dead end" jobs with poor pay and prospects or "stepping stones" to permanent employment in good jobs (Booth *et al.*, 2002). Gagliarducci (2005) selects a sample of individuals who enter the labour market via temporary employment and follow them until they obtain a permanent contract. He finds that the probability of moving from a temporary to a permanent job increases with the duration of the contract, but decreases with repeated temporary jobs and especially with interruptions. This suggests that it is not temporary employment per se but the intermittence associated with it that is detrimental to employment prospects. Güell and Petrongolo (2007), using a duration model with competing risks of terminating into permanent employment versus alternative states, find that conversion rates from temporary to permanent jobs increase with tenure. Van den Berg *et al.* (2004) investigates locking-in effects of temporary subsidized jobs using a natural experiment that occurred in the Slovak labour market in the early 1990s. He finds that if the subsidized job lasts too long, workers start reducing their job search intensity. In addition to this, the idea of looking at repeated temporary contracts is not a new phenomenon. Booth *et al.* (2002) study the effect of the number of temporary contracts held in the past on current wages. Zijl *et al.* (2004) take into consideration the presence of multiple spells for identification purposes, finding that temporary jobs serve as stepping-stones towards regular employment.

In general, national and international literature focus on the effect of temporary contracts against permanent contracts. A Permanent contract represents a target to achieve, the job of good quality. It is conceived as an absorption state, synonymous of stability, but, as previously stated, it does not always represent the final target to achieve. It is believed that a permanent contract assures more certainties, social protection (maternity, retirement, etc ...), stability and continuity in the work experience. On the other hand, contracts with temporary duration (fixed term contract, temporary contract, Co.Co.Pro, etc ...) are considered unstable because of their shorter duration and less security. Actually it is not always true that a permanent contract guarantees *a priori* a more stable work. As a fact, in some European countries (UK, Scandinavian countries, Switzerland, etc ...), a low social protection leads to a very high frequency of interruptions in permanent contracts, while in Italy we face a similar phenomenon due to the presence of a big amount of little firms (less than 15 employees) that often have a short life. Fabrizi (2009) analyzes the Italian workers' careers

from 1998 to 2004; he finds out that 11% of people that started working with a permanent job, after 1, 3 or 6 years had their contracts changed into more flexible ones. Furthermore, after 6 years from the entrance in the working world, only the 46% of workers are still under a permanent contract. At regional level, in Veneto, Accornero *et al.* (2000) shows that about 50% of permanent contracts ends during the first year. Similar results can be found in Lucidi and Raitano (2009). Of great interest is the province of Milan, characterized by a high labor mobility. Mezzanzanica (2008) analyses the contracts started from 2000 to 2007 and finds out that the average duration of the permanent contracts is about 19 months. This duration is longer than the average of temporary contracts, but not comparable with the duration of the permanent contracts of the previous decade.

4.1.4 Research hypothesis

This evidence underlines how permanent contracts *per se* are no more a meaning of stability and continuity. In this context it is better to define stability in terms of other factors, regardless of the type of contract (Fabrizi, 2009). Following this direction an important factor is the contract duration: stability has to be measured by the actual time worked and it has to be associated with a longer work duration. It is so necessity to define again which characteristics are peculiar of a good job. Considering the subsample of people that get a degree, it is interesting to take into account both work duration and work coherence related to one's own studies. To the best of my knowledge the majority of papers study the impact of the temporary (or fixed-term) job(s) on the career, measuring the time to get a permanent job (see for example Gagliarducci (2005); Bonnal *et al.* (1997); Gritz (1993)). The latter is usually considered the target job, but as explained above the permanent job is no more to be considered a point of arrival, given that its duration can be also very short. Usually after the first permanent contract other work experiences are observed. For this reason, with the data at hand it is particularly important to investigate what happens after getting the first long-term job. I call "stable" a job with a duration of at least 540 days (1 year and a half). This duration seems appropriate as it is longer than the ordinary duration of two successive fixed-term contracts. As the contract type is no more an appealing characteristic replaced by the effective duration, in the subsample of people with a degree, the coherence with the field of studies assumes a great relevance in order to define a good job. To the best of my knowledge in literature there are no studies that deal with the issue of work coherence. Scope of my work is indeed to study the impact of coherence of the first stable job (as defined above) on the following career. In this context lots of interesting questions arise. Can a coherent long experience lead the subject up to have another coherent experience? In other words, does a long enough experience with a certain coherence

have an impact on the future coherence experience? Which variables have a significant effect on the work coherence?

4.2 Data

The database for the analysis is provided by the Inter-University Research Center on Public Services (CRISP) holding the following administrative databases:

- the observatory of the Lombardy job market database supplying information relative to all the obligatory information given by the employer regarding public and private employees from January 1, 2000 to nowadays. For each subject it contains information regarding each individual work experience described by the date of entry, duration and cessation of employment, type of sector and qualification.

- the databases concerning graduates from three of the biggest universities of Milan reporting the marks received from individual exams and the final graduation status and title of every student that obtained the degree during the period between 2003 and 2008.

It is worth mentioning that access to databases of this nature is extremely rare, even at an international level, because data come from administrative sources and not from sample groups. Note that from the administrative archives, the employment status of a subject is not available if he is: not employed, employed outside the Lombardy region, self-employed, or employed in the public sector or with a coordinated and continued collaboration type of contract.

In particular these datasets contain the following information:

- working relationship (contract start and end date, extension or transformation date, contract type, qualification). With such information it is possible to construct the sequence of events that represents the worker career;

- worker (age, gender, domicile);

- worker's firm (economic sector);

- university career (faculty, course of study, graduation mark, time to get the degree and type of high school).

4.2.1 Sample considered

All the working relationships before the degree or with a duration of less than 20 days have been eliminated from the database as not representative

for the individual career. Contracts started before the degree and terminated after the degree have been considered. The variable "Coherence", showing if there is coherence between the job qualification and one's own field of studies, has been constructed comparing the course of study with the job qualification for each faculty. It is assigned value 1 when qualification resulted coherent with the one explicitly expressed from the website of each faculty, 0 otherwise. I chose the faculties for which the qualification were more clear and objective for the construction of the coherence variable. The faculties considered are:"Mathematical, Physical and Natural Sciences" (MPNS), "Economics", "Faculty of education", "Psychology" and "Social Sciences". In order to minimize the initial sample heterogeneity, have been selected only workers who, at the time of the degree, were aged between 21 and 35 (this restriction is common enough, see for example Gagliarducci (2005)). The subjects analyzed are restricted to those who got the degree between 2003 and 2005. In this way there are at least 4 years of work history after the degree to investigate on, given that the available data on their subsequent work history is up to July 2010. To take into account the students'"ability" the graduation mark is considered. A dichotomous variable is constructed considering as skilled those students that got a graduation mark equal to or higher than 106. This classification has been chosen because the class 106-110 is the last one of *ALMA LAUREA* classification.

To simplify the result interpretation some variables have been reclassified following the *ALMA LAUREA* classification:

- age at the degree reclassified in macro-classes: 20-21 years old,22, 23, 24, 25, 26, 27, 28, 29, 30-35 years old;
- type of high school reclassified in the few classes: secondary school focusing on sciences, secondary school focusing on humanities, training college, other.

After the cleaning procedure, the resulting subsample is formed by 94,464 working relationships referred to 25,871 people graduated in Milan and working in the Lombardy region.

Given that I want to study both stability and coherence I select two particular subsamples from it. In both subsamples the treated are represented by people for whom I observe at least a coherent stable job after the degree. I choose the first stable job observed. The two subsamples differ for the control group. In one of them I consider the subset of people for whom the longer work experience observed is coherent with one's own studies. In this case the control group can be considered as a kind of different treatment consisting in an unstable job with duration that can vary from 20 days to 539 days. With this subsample I can evaluate the effect of stability, given that the two groups (treated and untreated) differ only for this characteristic. Indeed the model used takes into account both observable and

unobservable variables that can affect both treatment and outcome. The second control group is formed by people for whom at least a stable job is observed, but the first one stable job is incoherent with one's own studies. With this subsample I can evaluate the effect of coherence, given that the two groups differ only for this characteristic.

The variable of interest is then the coherence associated with the work experiences subsequent the treatment. It assumes the folloing values:

0 if the work coherence is unknown (in this case the subject cannot be present in the dataset or the coherence value is missing, given that there is no information on qualification at that time);

1 if the job qualification is incoherent with the field of study;

2 if the job qualification is coherent with the field of study.

Note that categories are ordered. This variable is considered in different time istants: three months before the beginning of the reference job, six and nine months after the end of it. The reference job is for treated the first cohehrent stable job and for the two control groups considered, the coherent unstable job and the inchoerent stable job, respectively. The other information about the individual career is synthesized trough the variables indicating the number of preceding and subsequent contracts. The former represent also a proxy of the distance from the degree to the first coherent stable job (the treatment).

4.2.2 Descriptive statistics

The purpose of this study is to evaluate the effect of the first stable job, where a stable job is a working relationship lasting not less than 540 days (1 year and a half) representing the treatment. This duration seems appropriate as it is longer than the ordinary duration of two successive fixed-term contracts. For this reason I have to identify the treated individuals with at least one stable job, who are 18,430 related to 72,100 work relationships. Among these individuals only those having at least one work relationship after the stable job were selected. This condition is necessary to study the effect of the first stable job on careers. There are 11,012 individuals meeting this characteristic related to 38,431 work relationships. Among these, 3,995 individuals had a job before the first stable one while for the remaining 7,017 the stable job is the first after graduation. The control sample is represented by workers with no stable job after graduation, which are 7,441 related to 22,364 work relationships. The covariates considered are: gender, graduation age, graduation marks, type of high school (scientific, humanities, etc ...), graduation program type (bachelor degree, master degree, old system degree), faculty, number of work experiences preceding and following the reference job. Individuals with at least one of these variables missing

have been cut off as it is assumed the missing values are random (missing at random). To apply the model all combinations of those variables have been identified, associated with the respective number of individuals having these characteristics. The number of combinations is equal to 14,292 related to 17,643 individuals. If I consider only combinations of individuals related to more than one individual, the number of combinations falls down to 1,913 related to 5,264 individuals.

Table 4.1 gives descriptive statistics for the subsample which contains 5,264 individuals. The table shows that the number of graduates has in-

(a) Demographics variables

Variables	**%**
Gender	
Male	**31.3**
Female	68.6
Age	
21-22	6.6
23	13.8
24	**20.5**
25	**20.3**
26	13.4
27	8.4
28	5.4
29	4.0
30-35	7.2
Maturity	
Ss on sciences	28.5
Ss on humanities	14.5
Technical college	**32.7**
Other colleges	24.1

(b) Education variables

Variables	**%**
Faculty	
Economics	23.3
Psychology	12.3
Education	31.1
MPNS	23.4
Sociology	9.7
Type of degree	
Bachelor's degree	**42.9**
Master's degree	11.1
Degree	45.8
Grade	
Grade $\leq$ 106	68.3
Grade > 106	**31.6**
Year of degree	
2003	13.8
2004	18.0
2005	22.1
2006	22.7
2007	23.2

Table 4.1: Demographics and Education variables

creased from 2003 to 2007. This is also due to an intrinsic feature of the database, given that considered individuals come from the intersection of the database of graduates and workers. Lately the database of workers has become more comprehensive because it includes more contract types and therefore the match between the two databases has led to a higher number of individuals. Among them 68.6% of individuals are females. This is because, among the 5 right choices, 3 are female-dominated (Education, Psychology and Social Sciences, Table 4.5), representing more than 50% of the population. In fact, 31.1% of the individuals have a degree in Education,

	Bachelor's degree	**Master's degree**	**Degree**
Degree duration			
missing	26.8	45.7	48.9
2	4.1	**30.3**	0.2
3	**17.4**	**21.6**	0.8
4	**28.7**	1.9	2.8
5	**14.3**	0.2	**12.8**
6	5.6	0.0	**14.0**
>6	2.5	0.0	**19.8**

Table 4.2: Degree duration for type of degree

	Grade ≤ 106		**Grade>106**	
	% col	**% row**	**% col**	**% row**
Gender				
Male	36.2	78.9	20.8	21.0
Female	**63.7**	63.4	**79.1**	36.5
Type of degree				
Bachelor's degree	**51.5**	82.0	24.4	**17.9**
Master's degree	6.2	38.5	**21.6**	**61.4**
Degree	42.1	62.8	53.8	**37.1**

Table 4.3: Descriptive statistics stratified for grade

12.3% in Psychology, 9.7% in Sociology, 23.4% in MPNS and 23.3% in Economics. Unlike the expectations 32.7% of the graduates come from technical and tertiary college and only 14.5% from the secondary school focusing on humanities despite the fact that more than 50% of individuals have a degree in humanistic field.

Most individuals (68.0%) are between 23 and 26 years old. Only 20.4% of individuals graduate before turning 23 despite the fact that 42.9% of graduates have a three-year degree. This shows that not everyone can graduate on time. This is also shown in Table 4.2 on the degrees duration where it is clear that the mode duration for a three-year degree is four years, for a specialistics 2 years and for the old system degree it is 6 years. Graduated with a graduation mark higher than 106 represent 31.6% of the considered individuals. As shown in Table 4.3, with higher marks are those coming from a specialistic degree, getting higher grades than the ones coming from a three-year degree and the old system degree. As many as 61.4% of individuals coming from master's degrees scored a mark higher than 106, against 17.9% of the three-year degrees and 37.1% of the old system degrees.

As shown in Table 4.5 the faculties of Psychology, Sociology and Educa-

	Male	**Female**
Age		
21-23	5.5	**7.1**
23	12.1	**14.6**
24	17.4	**21.9**
25	19.7	**20.6**
26	**15.2**	12.5
27	**9.9**	7.8
28	**6.5**	4.9
29	**4.8**	3.6
30-35	**8.4**	6.7

Table 4.4: Distribution of age stratified for gender

	Economics		**Psychology**		**Education**		**MPNS**		**Sociology**	
	% col	**% row**	**% col**	**% row**	**% col**	**% row**	**% col**	**% row**	**% col**	**% row**
Gender										
Male	48.2	**35.8**	14.3	5.6	11.7	11.6	55.4	**41.4**	17.4	5.3
Female	**51.7**	17.6	**85.6**	15.4	**88.2**	40.0	**44.5**	15.2	**82.5**	11.6
Grade										
Grade ≤ 106	84.7	28.9	62.1	11.2	55.1	25.1	69.5	23.8	76.2	10.8
Grade>106	**15.2**	11.2	**37.8**	14.7	**44.8**	**44.0**	30.4	**22.5**	23.7	7.2
Course type										
Ss on sciences	40.3	33.0	39.0	16.9	23.3	25.4	19.1	15.7	25.8	8.7
Ss on humanities	5.9	9.5	18.4	15.6	9.1	19.4	**30.2**	**48.6**	10.1	6.7
Technical college	**47.2**	**33.7**	21.9	8.2	**34.3**	**32.6**	19.8	14.1	**37.6**	11.1
Other colleges	6.4	6.2	20.5	10.5	33.2	42.7	30.7	29.8	26.4	10.6
Type of degree										
Bachelor's degree	46.7	25.4	27.8	8.0	**36.8**	26.6	43.2	23.5	**72.3**	16.3
Master's degree	11.1	23.4	**23.2**	25.7	**4.6**	12.9	14.8	31.0	7.7	6.7
Degree	42.0	21.4	48.9	13.1	58.5	39.7	41.9	21.4	19.8	4.1

Table 4.5: Descriptive statistics stratified for faculty

tion have a percentage of females over 80%, while in the faculty of Economics and MPNS the percentage of females is 51.7% and 44.5%, respectively. The faculties for which I observe a greater number of graduates with higher marks than the average are Education, with 44.8% of "skilled", and Psychology, with 37.8%, while faculties with lower numbers of "skilled" graduates are Economics and Sociology. This may be mainly due to two factors. These particularly high marks can be explained by the fact that these faculties are those in which women are predominant. Females, in fact, tend to graduate sooner (see Table 4.4) and with higher marks than males (see Table 4.3). The lower marks for graduates in sociology may be due to the higher incidence of three-year degrees, equivalent to 72.3% (see Table 4.5), over the average values. In the average, in fact, 42.9% of individuals go for a three-year degree, 11.1% for the master's degree and 45.8% for the old system degree (see Table4.0(b)).

Work Variables	**%**
Modality of work	
Full time	63.2
Part-time	**30.5**
Undefined	6.1
Contract type	
Temporary contract	14.8
Fixed term contract	**54.1**
Permanent contract	**29.1**
Undefined	1.8
Level	
incoherent-medium level	**29.1**
coherent-medium level	5.3
incoherent-high level	10.4
coherent-high level	**41.6**
Undefined	13.4

Table 4.6: Work variables

Interesting results come out analyzing where the high school graduates come from. In contrast to expectations, 47.2% of graduates in Economics, 34.3% of graduates in Education and 37.6% of graduates in Sociology come from a technical or tertiary college. While 48.6% of individuals come from humanities graduate in MPNS.

I can group workers states into three main categories: Permanent contract (PC), Fixed-term contract (FC) and Temporary contract (TC). The PC state includes workers employed under a Permanent contract, people doing homework and apprenticeship. The Fixed-term contract includes workers employed under a Fixed-term contract and Co.co.co (a contract introduced in 1993 referred to particular project instead of a time-period). Temporary contract includes "lavoro interinale" and people with "contratto di inserimento" or training (a contract introduced in 1985 to provide people between 16 and 32 years old with training opportunities). The Table 4.6 shows that most of the contracts are Fixed-term (54.1%), while 29.1% are Permanent contracts and only 14.8% are Temporary contracts. Employment is spread between full time engagement, 63.2% and part-time 30.5%. The last one includes horizontal and vertical part-time. Table 4.7 shows the composition of the work category: most of permanent and temporary contract are full-time (respectively 75.7% and 68.7%), while most of part-time contracts are fixed-term (61%).

As expected males get more full-time and permanent jobs while females get more part-time and fixed term job (see Table 4.8). Figure 4.1 presents

	Full time		Part-time	
	% col	% row	% col	% row
Type of contract				
Permanent contract	34.8	75.7	22.7	23.8
Fixed term contract	47.0	54.9	61.0	34.4
Temporary contract	16.1	68.7	14.1	29.1
Undefined	1.8	65.8	2.0	34.1

Table 4.7: Type of work variable stratified for modality of work (undefined contract is omitted in the table)

Kaplan-Meier estimates for survival functions of durations stratified for each type of contract. There curves show that the exit rate from permanent job is lower than the exit rate from the temporary and fixed-term jobs. Looking at the box-plot in Figure 4.2 the conclusion is the same: the median duration of permanent contract is about 2 years (24 months), while the median duration of fixed-term and temporary contract is about 6 and 10 months, respectively. The median is used for comparison, because it is less affected by extreme values.

The interquartile range underlies where 50% of the values fall: for temporary contracts the interquartile range is about 2-18 months, that becames wider for fixed-term contracts, 4-22 months, and much wider for permanent contracts, 10-42 months. Besides that, permament contracts have a bigger standard deviation suggesting a bigger variability compared with other type of contract. The median value is not as high as to be defined a "permanent" contract. Furthermore the minimum value assumed by this type of contract is 20 days underlying a "short" duration for such contract. In most studies PC is considered an absorption state, meaning that every spell after the

	Male	Female
Work modality		
Full time	**73.1**	58.7
Part-time	20.4	**35.1**
Undefined	6.4	6.0
Type of contract		
Permanent contract	**35.4**	26.2
Fixed term contract	44.1	**58.7**
Temporary contract	**18.8**	13.0
Undefined	1.5	1.9

Table 4.8: Work variable stratified for gender

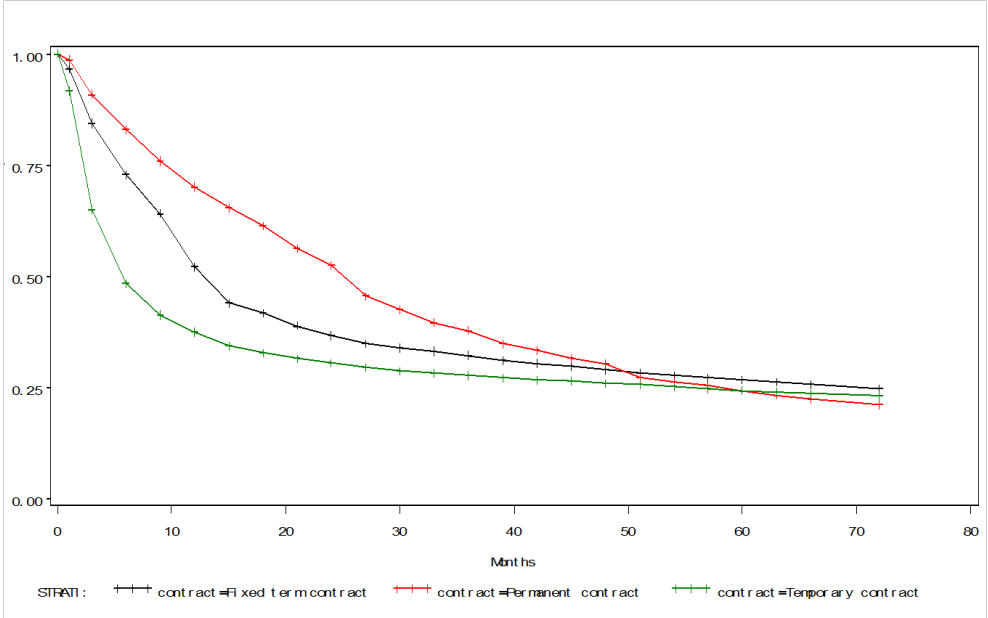

Figure 4.1: Duration - K-M Estimates

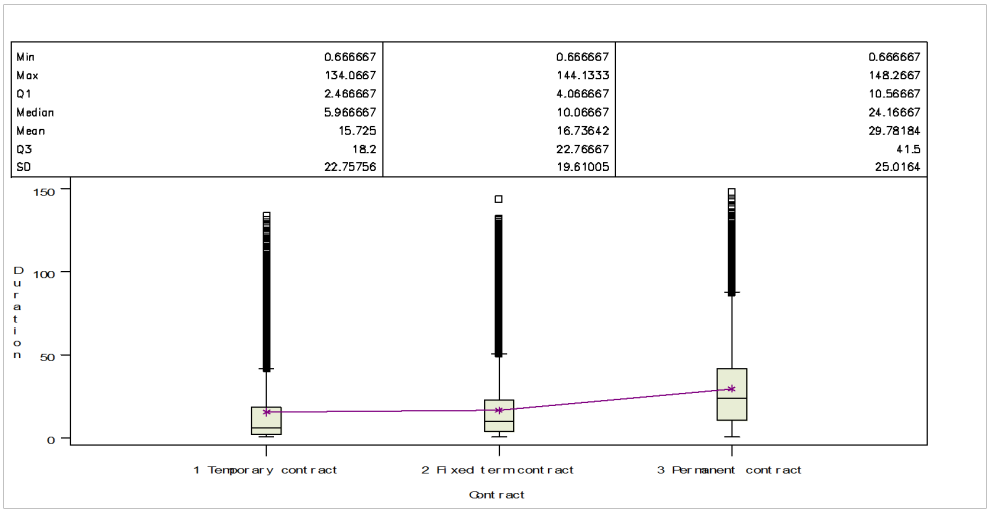

Figure 4.2: Boxplot per type of contract (duration expressed in months)

transition to a PC does not have to be considered as it is quite occasional. This analysis shows that it is not true. The high presence of "outliers" in the box-plot in Figure 4.2 underlies the big right asymmetry of duration distributions: there are many contracts with a short duration and few contracts with a long duration. As explained above, the goodness of a job cannot be described only by the type of contract. A good choice needs to consider the real duration of a job, the coherence and the qualification level of the considered job. Table 4.6 shows that 41.6% of the jobs are coherent with the field of study and of high qualification suggesting that a high percentage of graduated people gets a good job. On the other hand, there is also a high percentage of jobs incoherent with the studies done and with a low-medium qualification (29.1%).

	Incoherent	**Coherent**
21-22	**8.1**	6.6
23	**15.8**	14.8
24	20.1	**22.4**
25	19.4	**21.4**
26	13.0	12.4
27	8.0	8.0
28	5.3	5.5
29	3.7	3.1
30-35	6.1	5.4

Table 4.9: Incoherent variable stratified for age

Descriptive statistics in Table 4.9 show the linkage between job coherence and age: worse jobs (incoherent) are concentrated more for younger graduates and better jobs (coherence) for older graduates.

It's worth noting the characteristics of the group for which can be observed a job longer than 540 days without the ending date: the censored group. This group can be considered the most stable as I do not know when and if the work relationship will end. Comparing the characteristics of this group with the average characteristics of the population, it can be noted that there are more males and people coming from scientific faculties (Economics and MPNS). There are, also, more people with grades below 106 and a higher number of students coming from scientific high schools and technical institutes.

Subsamples considered: main differences with the whole sample

As explained before, I considered two subsamples of this population. The first subpopulation is represented by people for whom the longer job experi-

ence represents a coherent job. In such a case the "treatment" is represented by having or not a stable job after the degree in the observation period. Through this subsample I want to verify if a stable coherence work favours a future coherence work compared with a non stable coherence work. The second subpopulation is represented by people that have at least a stable job: in this case the "treatment" is represented by the coherence of the stable job.

The first subsample is made up of 1,966 people. This differs from the entire sample because there are more males (32.7% vs. 31.3%) and therefore more people from science faculties. 27.6% (vs. 23.3%) come from Economics and 25.6% from Mathematical, Physical and Natural Sciences (vs. 23.4%). There are more individuals who come from the old system (51% vs. 45.8%), but despite this, the age at graduation is lower. The percentage of individuals with high grades get up, pointing out that perhaps there is a correlation between skill and consistency in the future work. The treated group in this case consists of those individuals who have a stable job (coherent), accounting for 72.6% of the sample. In this case, the treated group is characterized, compared to controls, by more males, by more individuals graduated in Economics coming from high schools and technical schools. Among the treated there are fewer people who have received high grades. Among the controls 50% of the individuals got a degree in education.

The second subsample is formed by 2,538 people. It is characterized by a males percentage (34,2% vs 31.3%) slightly higher than the whole sample. There is no surprise, then, for a higher percentage of individuals studying Economics and coming from a scientific and technical high school. As shown in the previous tables, males are usually less clever than females, thus the percentage of individuals with a graduation mark higher than 106 gets low to 27% (vs 31,6%). Percentage of treated belonging to this subsample, having a (stable) coherent job then, is 55.1%. There is no big difference between treated group and the control group. Comparing the conditional distributions in the treatment group showed that the treated group compared with controls has a slightly higher number of males, higher number of individuals with high grades, fewer individuals from three-year degrees and more individuals from the Faculty of Mathematical, Physical and Natural Sciences.s 31,6%). Percentage of treated belonging to this subsample, having a (stable) coherent job then, is 55.1%. There is no big difference between treated group and the control group. Comparing the conditional distributions in the treatment group showed that the treated group compared with controls has a slightly higher number of males, higher number of individuals with high grades, fewer individuals from three-year degrees and more individuals from the Faculty of Mathematical, Physical and Natural Sciences.

Chapter 5

Effect of the first stable coherent job: a dynamic logit model approach

In this chapter I investigated the impact of a stable coherent work on the future job coherence according to the empirical data illustrated in the previous chapter. First I want to present the chosen dynamic causal model according to the nature of the response variable of interest. Such model allows to estimate in a simple and efficient way the dynamic causal effect of interest. It is characterized by the fact that the unobserved heterogeneity is modeled by a discrete latent variable to avoid parametric assumptions. The estimation method which is carried out through the EM algorithm is explained in, as well as the way to compute standard errors for the model parameters. The results are illustrated for the model fitted to the subsamples of subjects described at the end of the previous chapter. For both subsamples the treatment is represented by the first coherent stable job and the response variable is the future job coherence. The two subsamples differ for the control group. In the former the longer work experience observed after the degree coherent with the university studies is considered. In such way the effect studied is represented by stability, given that the two groups (treated and control) differ only for the work experience length. Stability has been defined in the previous chapter as a work experience of at least 540 days. In the latter the first stable experience after the degree incoherent with university studies is taken into consideration. In this case the focus is on the coherence effect.

5.1 The proposed model and its main assumptions

I am going to illustrate the chosen model for the analysis of the dynamic effect and the main assumption underlying it. Given that the variable of in-

terest coherence, has been defined as a categorical variable with more than two levels, the dynamic logit model presented in Section 3.3 is used. In particular, this model was used in Bartolucci and Pennoni (2010) in a similar context of labour market histories. The categorical response variable y_{it} denotes the coherence associated to the work experience that is going on in the instant t for the individual i, $i = 1, \ldots, n$ with $t = 0, \ldots, T-1$. Three temporal instants are considered ($T = 3$) to analyze coherence trend in time. For each individual i in the sample, I denoted by y_{i0} the coherent job observed three months before the beginning of the treatment. Where the treatment, as explained above, is the first stable job after the degree coherent with the university studies: I denoted it by z_i. I denoted by y_{i1} and y_{i2} the coherence job observed six and nine months after the end of this job. The coherence variable assumes value 0, if the coherence job of the individual is not observed in that particular instant, value 1, if the job is incoherent and value 2, if the job is coherent with one's own studies. It is important to remember that coherence job of the individual is not observed if the subject is not present in the dataset in that particular instant or if the qualification is unknown for the work experience considered. These two categories represent two distinct sets of information, but in such case they have been considered together given that the distinction is not particularly relevant to the interest of the study. Given the ordinal nature of the response variable a model based on nested logit is used (see Section 3.3). Furthermore, given the three ordered categories of the coherence variable two logits are constructed. The first one compares the probability of having a job with a known coherence against the probability of having a job with an unknown coherence. Thus comparing categories 0 against all the other categories. At nested level, a cumulative logit model for the conditional probability of each category larger than 0 is used. In this case there is only a nested logit. The latter logit compares the probability of having a coherent job some months after the treatment z_i (6 and 9 months) against all the other categories. It is the more interesting one, given that from it, it is possible to capture information on the characteristics that lead up to a coherent work compared with an incoherent one. The model takes into account the number of subjects that share a particular level of the different observed covariates. The model account for unobserved heterogeneity and state dependence by the inclusion of subjective-specific intercepts and the lagged response variables among the regressors. The unobserved heterogeneity is modeled by a discrete distribution with k point of supports. The points identify k latent classes in the population and are represented by the random intercepts (see also Section 3.1.1). The model considers the first response variable y_{i0} as given, whereas the distribution of y_{it}, $t = 1, 2$ is modeled through logits. In particular, y_{i1}

is modeled as follows:

$$\log \frac{p(y_{i1} > 0|c_i, \mathbf{x}_{i1}, y_{i0})}{p(y_{i1} = 0|c_i, \mathbf{x}_{i1}, y_{i0})} = \alpha_{1c_i} + \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \sum_{j=1}^{2} d_{ij0}\beta_{1,j+1}$$

$$\log \frac{p(y_{i1} > 1|c_i, \mathbf{x}_{i1}, y_{i0}, y_{i1} > 0)}{p(y_{i1} \leq 1|c_i, \mathbf{x}_{i1}, y_{i0}, y_{i1} > 0)} = \alpha_{2c_i} + \mathbf{x}'_{i1}\boldsymbol{\beta}_2 + \sum_{j=1}^{2} d_{ij0}\beta_{2,j+1}$$

where:

- $\mathbf{x}_{i1}$ is the vector of exogenous covariates at the first occasion ($t = 1$);
- c_i is the latent class of subject i;
- α_{1c_i}, α_{2c_i} are the subjective-specific intercepts that represent the support points associated to latent class c ($c = 1, \ldots, k$);
- $d_{ij,t-1}$ is a dummy variable equal to 1 if the lagged response variable $y_{i,t-1} = j$ and 0 otherwise.

The probability of each latent class is denoted by π_c. For what concerns the distribution of $y_{it}, t = 2$, it is assumed

$$\log \frac{p(y_{it} > 0|c_i, \mathbf{x}_{it}, y_{i,t-1}, z_i)}{p(y_{it} = 0|c_i, \mathbf{x}_{it}, y_{i,t-1}, z_i)} = \alpha_{1c_i} + \mathbf{x}'_{it}\boldsymbol{\beta}_1 + \sum_{j=1}^{2} d_{ij,t-1}\beta_{1,j+1} + z_i\gamma_{1t}$$

$$\log \frac{p(y_{it} > 1|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)}{p(y_{it} \leq 1|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)} = \alpha_{2c_i} + \mathbf{x}'_{it}\boldsymbol{\beta}_2 + \sum_{j=1}^{2} d_{ij,t-1}\beta_{2,j+1} + z_i\gamma_{2t} \tag{5.1}$$

where the vector $\mathbf{x}_{it}$ includes also time dummies. The parameters γ_{1t} and γ_{2t} ($t = 2$) measure the dynamic effect of the treatment for each period. They correspond to the logit difference of the probability of improving one's own career between subject receiving and not receiving the treatment while all other factors remain constant. Finally for the binary variable z_i, equal to 1 if subject i has a stable job and to 0 otherwise, it is assumed:

$$\log \frac{p(z_i = 1|c_i, \mathbf{x}_{i1}, y_{i0})}{p(z_i = 0|c_i, \mathbf{x}_{i1}, y_{i0})} = \alpha_{3c_i} + \mathbf{x}'_{i1}\boldsymbol{\delta}_1 + \sum_{j=1}^{2} d_{ij0}\delta_{j+1}$$

with α_{3c_i}, $c = 1, \ldots, k$ being the support points associated to the latent classes; δ_1 the parameters associated to the covariates that affect the probability to get the treatment and d_{ij0} the parameter associated to the initial period.

In the model presented above it is assumed that all observable factors (represented by the covariates) and unobservable factors (represented by the

random intercepts) affecting both the coherence job status and the choice of the treatment are properly taken into account. Furthermore taking the sample of people from the flow of entrants in a particular state (that is people that get the first stable job after the degree) also the NA (no anticipation) assumption is satisfied (Section 3.4). Indeed I can assume that potential outcomes are determined externally, and are not affected by subject actions in response to different predictions of future outcomes, given that people do not know if and when they will get a long-term job. Under such assumptions a causal model in the sense of Abbring and Van den Berg (2003b) results. Indeed, causal models for observational studies similar to the present one are typically formulated following a potential outcome approach as illustrated in Chapter 1. Here, the potential outcomes may be denoted by $y_{it}^{(1)}$ and $y_{it}^{(0)}$ and, for every subject i and time occasion t, indicate the job's coherence if the treatment was or was not verified. It is worth noting that the model presented above is equivalent to a model formulated on these potential outcomes through a similar parameterization. In a related context, the equivalence between the two formulations is derived in Bartolucci (2010) and Ten Have *et al.* (2003). The main assumption for this equivalence to hold is that the potential outcomes are conditionally independent of z_i given the observed covariates and the random intercepts. An important aspect is that the parameters γ_{ht} in Equation (5.1) may be seen as suitable contrasts, on the logit scale, between the probabilities of certain configurations of $y_{it}^{(1)}$ and $y_{it}^{(0)}$. This enforces their interpretation as causal parameters.

5.2 Estimation method through the maximum likelihood

Estimation of the parameters is based on the maximization of the log-likelihood by the EM algorithm. The log-likelihood is:

$$\ell(\boldsymbol{\theta}) = \sum_i log[p(y_{i1}, z_i, y_{i2}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, y_{i0})],$$

where $\boldsymbol{\theta}$ denotes the vector of all parameters. As usual, the algorithm alternates the E- and M- steps until convergence and it is based on the complete data log-likelihood. On the basis of the dummy variable u_{ic} the latter may be expressed as:

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_i \sum_c u_{ic} log[p(y_{i1}, z_i, y_{i2}|c, \mathbf{x}_{i1}, \mathbf{x}_{i2}, y_{i0})\pi_c] = \\ &= \sum_i \sum_c u_{ic} log[p(y_{i1}|c, \mathbf{x}_{i1}, y_{i0})] + \sum_i \sum_c u_{ic} log[p(z_i|c, \mathbf{x}_{i1}, y_{i0})] + \\ &+ \sum_i \sum_c u_{ic} log[p(y_{i2}|c, \mathbf{x}_{i2}, y_{i1}, z_i)] + \sum_i \sum_c u_{ic} log(\pi_c)\end{aligned} \quad (5.2)$$

where u_{ic} is equal to 1 if subject i belongs to latent class c and 0 otherwise. At the E-step, the EM algorithm computes the conditional expected value of u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, k$, given the observed data and the current value of the parameters. This expected value is denoted by $\hat{u}_{ic}$ and is proportional to:

$$p(y_{i1}, z_i, y_{i2}|c, \mathbf{x}_{i1}, \mathbf{x}_{i2}, y_{i0})\pi_c.$$

The M-step consists of maximizing the expected value of the complete data log-likelihood, obtained by substituting in Equation (5.2) each u_{ic} by the corresponding expected value computed as above. In this way the parameter estimates are updated. In particular, to update the probabilities of the latent class there is an explicit solution given by $\pi_c = \sum_i \hat{u}_{ic}/n$, $c = 1, \ldots, k$. For the other parameters an algorithm to maximize the weighted log-likelihood of a logistic model is needed. A crucial point is the initialization of the EM algorithm. Different strategies may be used in order to overcome the problem of multimodality of the likelihood. As usual, it is convenient to use both deterministic and stochastic rules to choose the starting values and to take, as maximum likelihood estimate of the parameters, $\hat{\boldsymbol{\theta}}$, the solution that at convergence corresponds to the highest value of $\ell(\boldsymbol{\theta})$.

5.2.1 The EM algorithm

The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing MLE in situation where, for the absence of some additional data, ML estimation would be straightforward. The name EM comes from the 2 steps of the algorithm: an expectation step (E-step) followed by a maximization step (M-step) (Dempster *et al.*, 1977). The Expectation-Maximization (EM) algorithm is a general approach to iterative computation of maximum likelihood (ML) estimates when algorithms such as the Newton-Raphson method may turn out to be more complicated. In particular the EM is used when the observations can be viewed as incomplete data. Although a problem is not an incomplete-data one, there may be much to be gained computationally by artificially formulating it as such to facilitate ML estimation.

The actual observed data Y are considered as a subset of some not fully observable complete data (Y, X). It is assumed the existence of a probability density function $f(x, y; \theta) = L(\theta)$ for the the joint distribution of X and Y given a parameter value θ. Let f be a probability density function only when considered a function of both x and y. While for a fixed value of y, f is a positive integrable function. The observed data y are a realization of the sample space $\mathcal{Y}$. The corresponding x in the sample space $\mathcal{X}$ is not observed directly, but only indirectly through y. It is assumed a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$ and that x is known only to lie in $\mathcal{X}(y)$, the subset

of $\mathcal{X}$ determined by the equation $y = y(x)$. The likelihood of interest $L_c(\theta)$ is written as the marginal of the higher dimensional function $f(x, y|\theta)$:

$$L_c(\theta) = g(y|\theta) = \int_{\mathcal{X}(y)} f(x|\theta)dx$$

The EM algorithm finds the value of θ that maximizes $L(\theta)$ given an observed y using the associated family $L_c(\theta)$. Then given the incomplete data specification, there are many possible specifications of the complete data function $L_c(\theta)$ that will generate $L(\theta)$. The estimate of θ is obtained solving the incomplete-data likelihood equation:

$$\frac{\delta logL(\theta)}{\delta\theta}$$

by proceeding iteratively in terms of the complete data log-likelihood function, $logL_c(\theta)$. As it is unobservable, it is replaced by its conditional expectation given y using the current fit of θ. Let θ_0 be some initial value of θ. Then on the first iteration, the E-step requires the calculation of:

$$Q(\theta, \theta_0) = E_{\theta_0}[logL_c(\theta)|y] = E_{\theta_0}[logf(\mathbf{x}, \theta)|y]$$

The M-step requires the maximization of $Q(\theta, \theta_0)$ in respect of θ over the parameter space Θ. A way θ_1 is estimated so that:

$$Q(\theta_1, \theta_0) \geq Q(\theta, \theta_0)$$

for all $\theta \in \Theta$. The E- and M- steps are carried out again, but with θ_0 replaced by the current fit θ_1. On the $(k+1)$-th iteration, the E- and M-step are defined as follows:

E-Step Calculate

$$Q(\theta, \theta_k) = E_{\theta_0}[logL_c(\theta)|\mathbf{y}]$$

M-Step Choose θ_{k+1} to be any value of $\theta \in \Theta$ that maximizes $Q(\theta, \theta_k)$, that is

$$Q(\theta_{k+1}, \theta_k) \geq Q(\theta, \theta_k)$$

for all $\theta \in \Theta$.

The E- and M- steps are alternated repeatedly until the difference:

$$L(\theta_{k+1}) - L(\theta_k)$$

changes by an arbitrary small amount. Dempster *et al.* (1977) show that the incomplete likelihood function $L(\theta)$ is not decreased after an EM iteration, that means the EM is a monotone optimization algorithm:

$$L(\theta_{k+1}) \geq L(\theta_k)$$

Hence convergence must be obtained with a sequence of likelihood values that are bounded above. It is not necessary to specify the exact mapping from $\mathcal{X}$ and $\mathcal{Y}$. All that is necessary is the specification of the complete data vector x and the conditional density of X given the observed vector y (the latter is needed to carry out the E-step). As the choice of complete data vector x is not unique, it is chosen for computational convenience in respect of carrying out the E-and M- step. When the complete data $f(x, \theta)$ come from an exponential family $Q(\theta, \theta_k)$, it is sufficiently simple to permit to compute the $Q(\theta, \theta_k)$ at a reasonable computational cost and to allow a closed-from maximization. It is possible to prove that the observed log-likelihood has increased after any EM step and that the algorithm converges to a local maximum of this function. However, this local maximum cannot be guaranteed to correspond to the global maximum since the likelihood may be multimodal. As usual, this problem may be addressed by trying different initializations of the algorithm and then choosing the parameter value which at convergence gives the highest value of likelihood. In order to increase the chance that the point at convergence is the global maximum, it is necessary to properly initialize the algorithm. Typically, a multiple-try strategy is adopted, which is based on combining a deterministic rule with one or more random rules. The first is a simple rule, which leads to a reasonable guess of the parameters obtained by fitting a simplified version of the adopted model. Then, as maximum likelihood estimate of the parameters it is taken the value corresponding to the highest log-likelihood at convergence of the EM algorithm. However, some simulation studies show that the chance of there being more than one local maximum is usually low when the number of observations is large in comparison with the number of parameters and the model assumed holds. The EM algorithm is very popular for the following reasons:

- it is very simple to implement;
- the M-step equations are so simple that they can be solved even for parameters that are subject to constraints;
- it is parametrization independent. Because the M-step is defined by a maximization operation, it is independent of the way the parameters are represented. Thus any invertible transformation of the parameter vector θ leaves the EM recursion unchanged.

Although the EM algorithm has been successfully applied in a variety of contexts, there are two issues that have led to some criticism. The First concerns the fact that in certain situations its convergence can be quite slow. This has resulted in the development of modified versions of the algorithm as well as many simulation-based methods and other extensions of it (see Mc Lachlan Krishana for extension, the property of the algorithm and propriety

for convergence). However methods to accelerate the EM algorithm do tend to sacrifice the simplicity ans stability proper of this algorithm. The second issue concerns provision of standard errors. The EM algorithm does not automatically provide an estimate of the covariance matrix of the MLE, as do some other methods, such as Newton-type methods. In the following paragraph different methods are proposed to estimate them.

5.3 Goodness of fit

5.3.1 Model selection

An important phase consists in the model selection. Any model is a simplification of reality. It should be, on one hand complex enough to fit the data well and, on the other hand, it should be simple to interpret, smoothing the data rather than overfitting them. A simple model that fits adequately has the advantages of model parsimony. If a model has relatively little bias, describing reality well, it tends to provide more accurate estimates of the quantities of interest. Other criteria besides significance tests can help to select a good model in terms of estimating quantities of interest. Usually the most used criteria are the Bayesian Information Criteria or BIC (Schwarz, 1978) and the Akaike Information Criteria or AIC (Akaike, 1974). Akaike's Information Criterion (AIC) is defined as:

$$AIC = -2 \times \ell(\boldsymbol{\theta}) + 2 \times p,$$

where $\ell(\cdot)$ denotes the maximum log-likelihood, $\boldsymbol{\theta}$ the vector of parameters, p the number of estimated parameters in the log-likelihood. Increasing the number of free parameters to be estimated improves the goodness of fit, regardless of the number of parameters in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting. The preferred model is the one with the lowest AIC value. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters. The Bayesian Information Criterion (BIC) is defined as:

$$BIC = -2 \times \ell(\boldsymbol{\theta}) + p \times \log(n)$$

where $\ell(\cdot)$ denotes the maximum log-likelihood, $\boldsymbol{\theta}$ the vector of parameters, p the number of estimated parameters and n the sample size. The Bayesian Information Criterion is also known as the Schwartz criterion. From Bayesian's point of view, this is an approximation of the integrated likelihood; from the likelihood point of view, this is a negative function of log-likelihood at convergence with a penalty of the number of parameters as a function of the sample size. Consistent with the concept that larger log-likelihood value

is better, smaller BIC value is better. In BIC, the penalty for additional parameters is stronger than the AIC one.

In estimating the model, the number of mass points is decided in accordance to the BIC value. As the number of parameters increases, the BIC value decreases until it reaches the lowest level, then it starts to increase. In practice, let us start by estimating a model with the lowest possible number of mass points and then add one mass point each time, until the BIC value starts to increase (Kamakura and Russell, 1989). To select the best model a lot of possible criteria can be used. A discussion about the different criteria can be found in Gaure and Røed (2007).

5.3.2 Testing parameter

As suggested by the recent statistical literature, see among others Agresti (1990), there are mainly three standard ways to use the likelihood function to test the significance of particular explanatory variables in a statistical model. In the following one I briefly reviewed these methods to test parameters: the Wald test, likelihood ratio test and the score statistics. The former is the one used in my application.

The effects are typically tested in a familiar way, by creating a ratio of the estimate to the estimate of the standard error:

$$\frac{\hat{\theta}}{SE(\hat{\theta})}.$$

The usual null hypothesis test is whether the coefficient is significantly different from zero. ($\theta = \theta_0 = 0$). This kind of ratio is usually distributed as a z or t. If significance is determined by the normal curve then z-test is often referred to as a *Wald's test statistic* (Wald, 1943). It consists in the ratio of the parameter and his standard error (SE):

$$\frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}.$$

This test has an approximate standard normal null distribution Z, so one refers to the standard normal table to obtain one or two side p-values. Equivalently, for the two side alternatives, z^2 has a chi-squared null distribution and the p-value is then the right-tailed chi-squared probability above the observed value.

A second method is the *likelihood ratio test*. It used the likelihood function through the ratio of two maximizations: L_0, maximization of the likelihood function without the variable (under the null hypothesis) and, L_1, maximization of the likelihood function with the variable. The ratio L_0/L_1 of these two likelihood cannot exceed one. The latter is always at

least as large as the former, since the former results from maximizing over a restricted set of parameter values. (Wilks, 1935) showed that:

$$-2\log\frac{L_0}{L_1} = -2(\ell_0 - \ell_1), \tag{5.3}$$

where ℓ_0 and ℓ_1 denote the maximized log-likelihood function. Using minus twice its log is necessary to obtain a quantity whose distribution is known and can therefore be used for hypothesis testing. The likelihood ratio test can be seen again through the concept of *deviance*. The deviance (D) is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the alternative that the more general saturated model holds:

$$D = -2\log\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right].$$

For purposes of assigning the significance of an independent variable the value of D with and without the independent variable is compared:

$$D(\text{model without the variable}) - D(\text{model with the variable}).$$

This represents the change in D due to the inclusion of the independent variable. Because the likelihood of the saturated model is common to both values of D being differenced then the likelihood ratio test explained above (Equation 5.3) is found. The likelihood ratio test compares the deviance of two models by subtracting the smaller deviance (model with more parameters) from the larger deviance (model with larger deviance). The difference is a chi-square test with the number of degrees of freedom equal to the number of different parameters in the two models. Test of a single parameter using the likelihood ratio test is asymptotically equivalent to the Wald test (p-values should be halved in each case).

The third method uses the *score statistics* (Fisher and Rao) that is the ratio of the score function evaluated in θ_0 ($H_0 : \theta = \theta_0$) to its null SE. This test is based on the slope of the expected curvature of the log-likelihood function $\log L(\theta)$ at the null value θ_0, since the score function is $S(\theta) = \delta \log L(\theta)/\delta\theta$. The score value tends to be larger in absolute value when the estimation of the parameter θ is farther from the null hypothesis ($\theta = \theta_0 = 0$). The variance of the score function can be individuated by the Fisher Information:

$$Var[S(\theta)] = E[S(\theta)^2] - E[S] = E[S(\theta)^2] =$$

$$= E\left[\left(\frac{\delta logL(\theta)}{\delta\theta}\right)^2\right] = -E\left[\frac{\delta^2 logL(\theta)}{\delta\theta}\right] = i(\theta)$$

evaluated at θ_0. The standard error is $[i(\theta_0)^{1/2}]$. Since the score statistics has an approximate standard normal null distribution, then the chi-squared

form of the score statistics can be written as follows:

$$\frac{[S(\theta_0)]^2}{i(\theta_0)} = \frac{\left[\frac{\delta \log L(\theta)}{\delta\theta}\right]^2}{-E\left[\frac{\delta^2 \log L(\theta)}{\delta\theta}\right]}$$

where the partial derivatives are in respect to θ and evaluated at θ_0. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log-likelihood compared with θ and the inverse information matrix, both evaluated at the H_0 estimates.

5.3.3 Standard Error

In the context of the model presented above a central problem is how to compute the standard errors for the parameters of the model. As it is known, one major shortcoming of the EM is that the observed information matrix is not obtained as by-product of the algorithm, which is useful to get an estimate of the precision of the estimated parameters. A basic result due to Louis (1982) is that if the complete data have distribution in a regular exponential family, the second derivative of the log-likelihood of the observed data can be expressed entirely in terms of the complete data log-likelihood or covariance matrix of the estimated parameters. Thus, the final point concerns how to compute the information matrix. For this aim, several methods have been proposed in the literature which exploit the results of the EM algorithm (McLachlan and Peel, 2000). Usually the standard errors of the estimates of the parameters are obtained approximating the covariance matrix by the inverse of the *observed information matrix* $I(\hat{\theta}, y)$, which is defined as:

$$I(\hat{\theta}, y) = -\left(\frac{\delta}{\delta\hat{\theta}}\ell(\hat{\theta}, y)\right)^2.$$

One way to proceed is to directly evaluate $I(\hat{\theta}, y)$ after the computation of the MLE $\hat{\theta}$. However, analytical evaluation of the second order derivatives of the log-likelihood may be difficult. To avoid this problem it is possible to compute the observed information matrix for the incomplete data problem in terms of the conditional moments of the score of the complete data log-likelihood function, which is obtained as by-product of the EM algorithm. The observed information matrix for the incomplete data problem $I(\theta, y)$ can be expressed in the form:

$$\begin{aligned} I(\theta, y) &= I_c(\theta, y) - cov_\theta[S_c(Y_c, \theta)|y] = \\ &= I_c(\theta, y) - E_\theta[S_c(Y_c, \theta)S_c^T(Y_c, \theta)|y] + S(y, \theta)S^T(y, \theta), \end{aligned} \tag{5.4}$$

where $S(Y,\theta)$ and $S_c(Y_c,\theta)$ denote the incomplete-data and complete-data score statistics, respectively. Furthermore it can be shown that the latter can be written as function of the former:

$$S_c(Y,\theta) = E_\theta\left[\, S(Y,\theta)|\, y\right] = E_\theta\left[\left.\frac{\delta log L_c(\theta)}{\delta\theta}\right| y\right].$$

where $L_c(\theta)$ denotes the complete-data log-likelihood function. The observed information matrix $I(\hat{\theta})$ can be computed as:

$$I(\hat{\theta}) = I_c(\hat{\theta}, y) - E_\theta[S_c(Y_c,\theta)S_c^T(Y_c,\theta)|y]_{\theta=\hat{\theta}}$$

since the last term in Equation (5.4) is zero as $\hat{\theta}$ satisfied $S(y,\theta) = 0$. Another information type method is based on the result that, in large sample from regular model for which the log-likelihood is quadratic in the parameters, the likelihood ratio test and Wald's test for the significance of individual parameters are equivalent. This means that the deviance change on omitting the variable is equal to the square root of the ratio of the parameter estimates to the standard error:

$$D(\text{model without the variable}) - D(\text{model with the variable}) = \left[\frac{\hat{\theta}-\theta_0}{SE(\hat{\theta})}\right]^2.$$

The latter is equal to the square of the t-statistics. Thus the standard errors can be calculated as the absolute value of the parameter estimate divided by the square root of the deviance change:

$$\text{SE}(\hat{\theta}) = \frac{|\theta|}{[D(\text{model without the variable}) - D(\text{model with the variable})]^{1/2}}.$$

This requires the fitting of a set of reduced models in which each variable is omitted from the final version of the model. If the standard errors found by an information-based approach are too unstable a bootstrap approach can be used (Efron, 1981).

5.4 Results

The estimation procedure of the model consists in maximizing the log-likelihood in respect to all the model and heterogeneity parameters repeatedly for alternative values of the number of latent classes (Heckman and Singer, 1984). Let us start out with one latent class or support point of the distribution (i.e. assuming absence of heterogeneity), and then expand the model with new latent classes until the likelihood is not able to increase any further (Gaure and Røed, 2007). The number of latent classes is chosen according to the maximum likelihood which cannot be made any larger by adding additional classes to the heterogeneity distribution. However to

choose the best model there are many information criteria that "punish" parameter abundance as explained in the paragraph above. The one used here is BIC. In this section results about the model are described for the two subsamples presented above. In the following paragraph the subsample of people who have the first stable job (the treated) or the longer job (the control) coherent with one's own studies are considered. While in the subsequent one the subsample of people who have the first stable job coherent with one's own studies (the treated) are compared with people that have a first stable job incoherent (control) with one's own studies. In order to simplify the reading of parameters associated to each category of each variable considered in the model, categories and category of reference in the logit are shown in Table 5.1.

Variable	**parameters**	**categories**	**category of reference**
gender	$\beta_{\cdot,2}$	female	male
graduation mark	$\beta_{\cdot,4}$	graduation mark ≥ 106	graduation mark < 106
high school	$\beta_{\cdot,5}$	Ss on humanities	Ss on sciences
	$\beta_{\cdot,6}$	Technical college	
	$\beta_{\cdot,7}$	Other colleges	
type of degree	$\beta_{\cdot,8}$	Master's degree	Bachelor's degree
	$\beta_{\cdot,9}$	Degree	
faculty	$\beta_{\cdot,10}$	Psychology	Economics
	$\beta_{\cdot,11}$	Education	
	$\beta_{\cdot,12}$	MPNS	
	$\beta_{\cdot,13}$	Psychology	
lag response	β_2	$y_{i,t-1=1}$	$y_{i,t-1=0}$
	β_3	$y_{i,t-1=2}$	$y_{i,t-1=0}$

Table 5.1: Categories of variables and category of reference in the logit model

5.4.1 Subsample of people with a coherent job

Given the first subsample and according to the BIC I chose a model with two latent classes, given that the maximum log-likelihood and the BIC values for a model with one, two and three latent classes are the ones reported in Table 5.2. The number of parameters considered is 65. The parameter estimates, with the SE and the T-test for the chosen model are reported in Table 5.3. These estimates have been compared with estimates obtained considering models with different number of latent classes: the estimates don't change a lot and the results are very similar. The unobserved heterogeneity is modeled by two latent classes with estimates probabilities equal to 0.9036 and 0.0964, respectively. Given that two latent classes are considered, two random intercepts are estimated for each logit as shown in Table 5.3. Despite the high probability of belonging to the first latent class, the

k	$\ell(\boldsymbol{\theta})$	**BIC**
1	-2,923.6	6,170.5
2	-2,521.8	5,406.3
3	-2,542.8	5,487.8

Table 5.2: Log-likelihood and BIC values of the model with different number of classes in the subsample of coherent people

second one seems to have a greater impact on the response variable. In both logits the random intercept associated to the second latent class assumes a higher value. The most interesting aspect is the value of the estimates of the parameters γ_{ht}, which measure the dynamic impact of the coherent stable job (the treatment). In particular most is referred to the second logit. Given the structure of the coherence variable, it represents the comparison between people with a coherent job against people with an incoherent job. These estimates indicate that the treatment in the second logit has a significant effect 9 months after the end of the treatment considered. This means that people who have a coherent long term ("stable") job have a high probability to have another coherent job after 9 months. For what concerns the parameters measuring the effect of the individual covariates on the response variable, males have a greater probability to get a coherent work (given that gender is a dummy variable that assumes value 1 for females and 0 for males). From Table 5.3 it comes out also that graduation mark has a positive significant impact on job's coherence (given that this variable assumes value 1 for graduation mark higher than 106 and 0 otherwise). Having a Master's degree or a degree of the old school system has a positive impact on future job coherence. Furthermore the faculties associated to a higher probability to have a coherent stable job in the future are the faculties of Education and Economics. An high number of past experiences has a negative impact on the probability to get a coherent job. The higher the number of not stable (shorter then 540 days) experiences before the first stable job, the higher the probability to get a incoherent job. Having too short experiences is not good for coherence. Gagliarducci (2005) has already shown, in a more general context, that having lots of temporary jobs has negative effects on the future occupation. Instead age and high school education do not have a significant impact on the future work coherence. A strong state dependence is also observed since all the parameters associated to the lagged responses are highly significant, indicating a strong persistence on the working coherence.

From results in Table 5.4 it emerges that covariates that have a significant effect on propensity to have a coherent stable job are gender, age at the degree, high school education, type of degree, number of past experiences.

First logit					
Effects		**estimates**	**s.e.**	***t*-statistic**	***p*-value**
intercepts	α_{11}	3.131	-	-	-
	α_{12}	30.452	-	-	-
time dummies	$\beta_{1,1}$	0.708	0.108	6.533	0.000
gender	$\beta_{1,2}$	1.170	0.347	3.372	0.001
age	$\beta_{1,3}$	-0.981	0.149	-6.568	0.000
graduation mark	$\beta_{1,4}$	-0.023	0.034	-0.693	0.488
high school	$\beta_{1,5}$	-0.820	0.129	-6.380	0.000
	$\beta_{1,6}$	1.422	0.169	8.405	0.000
	$\beta_{1,7}$	0.513	0.128	4.010	0.000
type of degree	$\beta_{1,8}$	2.729	0.153	17.804	0.000
	$\beta_{1,9}$	-0.562	0.218	-2.579	0.010
faculty	$\beta_{1,10}$	1.456	0.139	10.507	0.000
	$\beta_{1,11}$	0.552	0.228	2.418	0.016
	$\beta_{1,12}$	2.309	0.167	13.858	0.000
	$\beta_{1,13}$	0.085	0.167	0.508	0.612
number of past experiences	$\beta_{1,14}$	-2.099	0.289	-7.265	0.000
lag response	β_{12}	7.750	0.082	94.629	0.000
	β_{13}	-0.341	0.152	-2.241	0.025
training	γ_{12}	-0.843	0.169	-4.980	0.000
Second Logit					
intercepts	α_{21}	-1.765	-	-	-
	α_{22}	2.211	-	-	-
time dummies	$\beta_{2,1}$	-6.013	1.464	-4.106	0.000
gender	$\beta_{2,2}$	-1.736	1.027	-1.691	0.091
age	$\beta_{2,3}$	-1.099	0.934	-1.176	0.239
graduation mark	$\beta_{2,4}$	0.059	0.017	3.471	0.001
high school	$\beta_{2,5}$	0.170	0.353	0.480	0.631
	$\beta_{2,6}$	-0.681	0.435	-1.567	0.117
	$\beta_{2,7}$	-0.004	0.298	-0.012	0.991
type of degree	$\beta_{2,8}$	1.281	0.432	2.966	0.003
	$\beta_{2,9}$	2.307	0.956	2.412	0.016
faculty	$\beta_{2,10}$	-0.117	0.324	-0.362	0.717
	$\beta_{2,11}$	1.061	0.505	2.103	0.035
	$\beta_{2,12}$	-0.929	0.401	-2.315	0.021
	$\beta_{2,13}$	-1.420	0.513	-2.766	0.006
number of past experiences	$\beta_{2,14}$	-0.290	0.039	-7.377	0.000
lag response	β_{22}	1.519	0.230	6.620	0.000
	β_{23}	0.018	0.007	2.586	0.010
training	γ_{22}	2.976	0.761	3.910	0.000

Table 5.3: Estimates of the parameters for the conditional probability of the response variable given the latent variables in the subsample of coherent people (†minus average age)

In particular, females have a higher propensity to have coherent stable job, as well as subjects with higher graduation marks. People who got a degree at a younger age and attended secondary school on humanities, have a higher propensity score to get the treatment. Furthermore also the Bachelor's degree seems to have a positive impact on coherence against degree of the new system. Finally the more favorable faculty for the job coherence seems to be Economics.

Effects		**estimates**	**s.e.**	***t*-statistic**	***p*-value**
intercepts	α_{31}	1.674	-	-	-
	α_{32}	0.341	-	-	-
gender	$\delta_{1,1}$	0.389	0.049	7.939	0.000
age	$\delta_{1,2}$	-1.286	0.143	-9.018	0.000
graduation mark	$\delta_{1,3}$	1.785	0.036	49.174	0.000
high school	$\delta_{1,4}$	-0.109	0.118	-0.922	0.356
	$\delta_{1,5}$	0.240	0.101	2.376	0.017
	$\delta_{1,6}$	0.040	0.127	0.316	0.752
type of degree	$\delta_{1,7}$	-0.667	0.143	-4.664	0.000
	$\delta_{1,8}$	-0.035	0.183	-0.191	0.848
faculty	$\delta_{1,9}$	-0.252	0.136	-1.849	0.064
	$\delta_{1,10}$	-0.472	0.210	-2.248	0.025
	$\delta_{1,11}$	-0.261	0.141	-1.858	0.063
	$\delta_{1,12}$	-0.544	0.160	-3.394	0.001
number of experiences	$\delta_{1,13}$	-0.670	0.313	-2.140	0.032
initial period	δ_2	-0.747	0.118	-6.321	0.000
	δ_3	0.799	0.154	5.185	0.000

Table 5.4: Estimates of the parameters for the conditional probability of having a stable job given the latent variables in the subsample of coherent people(†minus average age)

5.4.2 Subsample of people with a stable job

In this subsample people that have at least a stable job have been selected. In such a case the treatment is represented by the coherence stable job, while the control is represented by the stable but incoherent job. The estimates of these models in Table 5.6 represent the comparison between these two groups and show that having a long work experience incoherent with one's own studies can lead to give up a work coherent with one's own studies. Also in this case two latent classes are selected according to the BIC criteria. These estimates have been compared with estimates obtained considering models with different number of latent classes: the estimates don't change a lot and the results are very similar. The log-likelihood and the BIC values

according to the number of classes are reported in Table 5.5. The estimates probabilities of the two latent classes of the unobserved heterogeneity are 0.1076 and 0.8924, respectively. In this case the second latent class has the higher probability. From Table 5.6 arouses that the two latent classes have the same impact on the response variable in the first logit, while the second latent class has a lower impact compared to the other one in the second logit.

k	$\ell(\theta)$	**BIC**
1	-4,215.7	8,880.0
2	-3,533.0	7,596.2
3	-3,713.0	7,915.4

Table 5.5: Log-likelihood and BIC values of the model with different number of classes in the subsample of stable people

Among the estimates of the parameters γ_{ht}, measuring the dynamic impact of the treatment, the more significant parameters are the ones of the second logit. This suggests that the probability of having a coherent stable job after 9 months at the end of the treatment is greater for people who had a coherent stable job.

For what concerns the parameters measuring the effects of the covariates on the response variable, it is evident that males with higher graduation marks tend to have a greater probability to find a coherent stable job. Among the different types of high schools, the one focused on sciences seems to be better to get a coherent job than the one focused on humanities or the tertiary college. Bachelor's degree and faculties of Economics and Education are associated to a higher probability to get a coherent job. This is an awaited result given also the descriptive statistics. Also in this subsample, a lower number of experiences has a positive effect on the future job coherence and a strong state dependence is observed. There is no evidence of a significant effect of graduation age compared to the previous subsample. From Table 5.7 covariates that have a significant effect on the propensity to have a coherent job are gender, graduation age and graduation mark. In particular males have higher propensity, as well as younger subjects that get a degree with a higher graduation mark. The only significant parameters for the faculty and the type of degree are those associated to the faculty of MNPS and to the degree of the old school system. They have a positive impact on the propensity to get the treatment.

In conclusion I can state that this model allows to give a causal interpretation to the treatment parameter given that observed and unobserved variables are taken into account. The two models give pretty much the same results, showing that both coherence and stability have an effect on

First logit					
Effects		**estimates**	**s.e.**	***t*-statistic**	***p*-value**
intercepts	α_{11}	5.160	-	-	-
	α_{12}	5.517	-	-	-
time dummies	$\beta_{1,1}$	0.777	0.122	6.351	0.000
gender	$\beta_{1,2}$	0.621	0.253	2.452	0.014
age	$\beta_{1,3}$	-1.576	0.140	-11.274	0.000
graduation mark	$\beta_{1,4}$	0.018	0.034	0.523	0.601
high school	$\beta_{1,5}$	1.189	0.141	8.432	0.000
	$\beta_{1,6}$	0.314	0.185	1.702	0.089
	$\beta_{1,7}$	-1.180	0.128	-9.206	0.000
type of degree	$\beta_{1,8}$	-1.056	0.174	-6.078	0.000
	$\beta_{1,9}$	-0.119	0.230	-0.519	0.604
faculty	$\beta_{1,10}$	0.568	0.141	4.016	0.000
	$\beta_{1,11}$	-1.836	0.220	-8.354	0.000
	$\beta_{1,12}$	0.412	0.167	2.474	0.013
	$\beta_{1,13}$	-0.432	0.163	-2.650	0.008
number of past experiences	$\beta_{1,14}$	0.469	0.235	1.995	0.046
lag response	β_{12}	-9.552	0.102	-93.736	0.000
	β_{13}	-1.149	0.830	-1.384	0.166
training	γ_{12}	0.423	0.578	0.732	0.464
Second Logit					
intercepts	α_{21}	4.248	-	-	-
	α_{22}	-1.317	-	-	-
time dummies	$\beta_{2,1}$	-6.547	0.718	-9.123	0.000
gender	$\beta_{2,2}$	-2.365	0.725	-3.264	0.001
age	$\beta_{2,3}$	-0.028	0.168	-0.168	0.867
graduation mark	$\beta_{2,4}$	-0.112	0.039	-2.857	0.004
high school	$\beta_{2,5}$	-0.363	0.171	-2.127	0.033
	$\beta_{2,6}$	-1.180	0.526	-2.243	0.025
	$\beta_{2,7}$	-0.388	0.748	-0.518	0.604
type of degree	$\beta_{2,8}$	-1.612	0.212	-7.589	0.000
	$\beta_{2,9}$	-1.431	0.409	-3.503	0.000
faculty	$\beta_{2,10}$	-0.788	0.362	-2.177	0.029
	$\beta_{2,11}$	1.366	0.564	2.422	0.015
	$\beta_{2,12}$	-1.075	0.766	-1.403	0.161
	$\beta_{2,13}$	0.240	0.024	10.213	0.000
number of past experiences	$\beta_{2,14}$	-1.332	0.269	-4.959	0.000
lag response	β_{22}	-2.801	0.149	-18.799	0.000
	β_{23}	2.988	0.503	5.942	0.000
training	γ_{22}	0.877	0.314	2.792	0.005

Table 5.6: Estimates of the parameters for the conditional probability of the response variable given the latent variables in the subsample of stable people (†minus average age)

Effects		**estimates**	**s.e.**	***t*-statistic**	***p*-value**
intercepts	α_{31}	-1.557	-	-	-
	α_{32}	-0.721	-	-	-
gender	$\delta_{1,1}$	-2.046	0.337	-6.071	0.000
age	$\delta_{1,2}$	-0.524	0.145	-3.614	0.000
graduation mark	$\delta_{1,3}$	0.814	0.027	30.373	0.000
high school	$\delta_{1,4}$	-0.070	0.104	-0.680	0.496
	$\delta_{1,5}$	0.002	0.027	0.071	0.943
	$\delta_{1,6}$	0.196	0.109	1.789	0.074
type of degree	$\delta_{1,7}$	-0.053	0.144	-0.368	0.713
	$\delta_{1,8}$	0.209	0.099	2.115	0.034
faculty	$\delta_{1,9}$	0.236	0.132	1.794	0.073
	$\delta_{1,10}$	0.154	0.164	0.938	0.348
	$\delta_{1,11}$	0.460	0.108	4.275	0.000
	$\delta_{1,12}$	-0.040	0.166	-0.242	0.809
number of past experiences	$\delta_{1,13}$	0.134	0.122	1.104	0.270
initial period	δ_2	0.398	0.129	3.090	0.002
	δ_3	0.198	0.175	1.133	0.257

Table 5.7: Estimates of the parameters for the conditional probability of having a coherent job given the latent variables in the subsample of stable people (†minus average age)

the same direction. The models differ in the heterogeneity distribution. In the first model the first latent class has the higher probability, equal more or less to 90%. Moreover it has a lower impact on the response variable compared with the second latent class. In the second model the situation is reversed, it is the second latent class that has a higher probability of about 90%. In this case this latent class has the same impact on the response variable for the first logit and a lower impact compared with the other latent class for the second logit. From both models I can conclude that having a stable job coherent with one's own university degree has a positive causal effect on the future coherence job in the long-term period. Furthermore, the strong state dependence shows the significant impact of the past job coherence. Following the results from both models, the main features that seem to have a significant impact on coherence are the subject's ability, measured through the graduation mark, and the time distance from the degree, measured with the number of past experiences. As expected, faculties that lead to a coherent job with a higher probability are the ones of Economics and Education.

Conclusion

The recent theoretical literature on program evaluation has built on a combined features of earlier work in both the statistics and econometrics literature. Even if this literature starts from different perspectives, in both cases the same central problem is studied: evaluate the effects of the exposure of a set of units to a program or treatment on some outcome of interest. In the last years lots of discussion has been done about the econometric structural approach and the statistical program evaluation approach. Structural model makes the preferences and constraints explicit with given individual decisions, that rule interaction among agents and the sources of variability across agents. These features facilitate finding answers to more policy questions, absent in the program evaluation literature. In the statistical literature there is the absence of explicit model. Fewer assumptions in terms of exogeneity, functional form, exclusion and distributional assumptions than the standard structural estimation literature in econometrics are attractive features of this approach. The greater simplicity of estimation favours replicability, transparency and sensitivity analysis. Despite the recent advances in the structural literature, fully-specified structural models are often still hard to compute. Heckman in his last paper tries to reconcile these two kinds of literature. He recognizes that in some situations the parameters required to forecast particular policy modifications are represented by a combination of subsets of the structural parameters, which are much easier to identify. They require fewer and weaker assumptions that can be bring back to the modern statistical literature.

I applied a dynamic version of these models in the context of the labour market, given that I had administrative panel data at my disposal. Dynamic models have been recently proposed in literature to face the fact that a treatment or a policy may be evaluated dynamically on time. Furthermore these models allow to control for unobserved heterogeneity and to estimate state dependence. Having at disposal administrative panel data on both Lombardy labour market and records of the graduates of three biggest University of Milan, I use such models to study the impact of the first "stable" job coherent with the university studies on the future job coherence. To the best of my knowledge there are no papers that focus on job coherence. Moreover most articles focus on unemployment and temporary job and their

effects on time duration or on probability to get a permanent contract. The latter is considered a stable job, a point of arrival. However in the last decade the concept of job and work stability has changed. The rapid spread of temporary employment and the increased instability of the market has aroused a new concept of work: the work path. Work relationships have become more flexible and workers, even those employed on a permanent contract, can no longer expect to remain employed for a long period in the same company. In the last years, the average length of employment relationships has significantly been reduced with a consequent higher rate of total turnover in the labour market. Nowadays a permanent contract can have also a short duration, a duration shorter than a temporary contract. Consequently after a permanent job another work path is present. It is of great interest to study not only the path up to the permanent job, but also the path subsequent to it. In this context the necessity arises to define which characteristics are peculiar of a good job. For sure a long effective duration, independently from the specific contract, and in the subsample of graduate people, coherence with one's own studies.

According with the above remarks I use a new definition of stable job as: work experience with a duration of at least 540 days. I then consider the coherence associated to that work experience evaluating the impact of the future job coherence. The dataset at my disposal results from the joining of two database: the observatory of the Lombardy job market database from January 1, 2000 to nowadays and the database concerning graduates from three of the biggest universities of Milan during the period between 2003 and 2008. From this resulting dataset I consider in particular the sample of subjects younger than 35 years old at the degree and graduates from the faculty of Economics, Mathematical, Physical and Natural Sciences (MPNS), Education, Psychology and Social Sciences.

The scope of my study is to examine the impact of the first stable job coherent with the university studies on the future job coherence. Given the nature of the variable of interest, coherence, a dynamic logit causal model has been performed. Given that I want to study both stability and coherence I select two particular subsamples. In both subsamples the treated are represented by people for whom I observe at least a stable job after the degree. I choose the first stable job observed. The two subsamples differ for the control group. In one of them I consider the subset of people for whom the longer work experience observed is coherent with one's own studies. With this subsample I can evaluate the effect of stability, given that the two groups (treated and untreated) differ only for this characteristic. Indeed the model used takes into account both observable and unobservable variables that can affect both treatment and outcome. The second control group is formed by people for whom at least a stable job is observed, but the first one stable job is incoherent with one's own studies. With this subsample I can evaluate the effect of coherence, given that the two groups differ only

for this characteristic.

From the results obtained applying the model to the two sets of data I can conclude that a stable job coherent with one's own university degree has a positive causal effect on the future coherence job in the long-term period. Furthermore, the strong state dependence shows the significant impact of the past job coherence. The main features that seem to have a significant impact on coherence are the subject's ability, measured through the graduation mark, and the distance from the degree, measured with the number of past experiences. As expected, faculties that lead to a coherent job with a higher probability are the ones of Economics and Education.

Further development, given the type of information available, could consist in exploiting a more complicated model that takes into account job duration in a more explicit way. In such a case a duration or event-history model can be used. Here to summarize the career some temporal instants before and after the treatment have been taken into account. Therefore it is considered the job coherence associated to the work experience in progress in such instants. An alternative way to proceed is to focus on subsequent work experiences and analyze the duration and coherence associated to each of them. The resulting model will be more complicated compared with the one used here, but less information will be lost. Furthermore in my work I used coherence with the university studies to define a good job. A further development could be to consider also the qualification level (skill) associated to each work experience. The question that arises is if it is better to get a coherent job with a low qualification or an incoherent job with a high qualification. In such context the main variables to consider to define a good job become three: duration, coherence and skill. Finally more faculties and more universities can be considered. Comparing the same faculties for different universities also the university effect can be performed.

Bibliography

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, **97**, 284–292.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, **72**, 1–19.

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, **70**, 91–117.

Abbring, J. (2008). The event-history approach to program evaluation. *Advances in Econometrics*, **21**, 33–55.

Abbring, J. and Heckman, J. (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of Econometrics*, **6**, 5145–5303.

Abbring, J. and Heckman, J. (2008). Dynamic policy analysis. *The Econometrics of Panel Data*, pages 795–863.

Abbring, J. and Van den Berg, G. (2003a). The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 701–710.

Abbring, J. and Van den Berg, G. (2003b). The nonparametric identification of treatment effects in duration models. *Econometrica*, **71**, 1491–1517.

Abbring, J. and Van den Berg, G. (2005). Social experiments and instrumental variables with duration outcomes.

Abbring, J., Van den Berg, G., and Van Ours, J. (2005). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *Economic Journal - London-*, **115**, 602.

Accornero, A., Anastasia, B., Gambuzzo, M., Gualmini, E., and Rasera, M. (2000). Solo una grande giostra. *La diffusione del lavoro a tempo determinato, Milano: Franco Angeli*.

Agresti, A. (1990). *Categorical data analysis*, volume 5. Wiley Online Library.

Agresti, A. (2002). Categorical data analysis (2002). *Hoboken, New Jersey: John Wiley & Sons Inc*, pages 267–313.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE transactions on Automatic Control*, **19**, 716–723.

Angrist, J. and Pischke, J. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton Univ Pr.

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

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**, 648–660.

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

Bartolucci, F. (2010). On the conditional logistic estimator in two-arm experimental studies with non-compliance and before-after binary outcomes. *Statistics in medicine*, **29**, 1411–1429.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**, 816–831.

Bartolucci, F. and Pennoni, F. (2010). Impact evaluation of job training programs by a latent variable model, In: Ingrassia S., Rocci R., Vichi M. New Perspectives in Statistical Modelling and Data Analysis.

Bauer, T. and Bender, S. (2004). Technological change, organizational change, and job turnover. *Labour Economics*, **11**, 265–291.

Bonnal, L., Fougère, D., and Sérandon, A. (1997). Evaluating the impact of French employment policies on individual labour market histories. *The Review of Economic Studies*, **64**, 683–713.

Booth, A., Francesconi, M., and Frank, J. (2002). Temporary jobs: stepping stones or dead ends? *The Economic Journal*, **112**, F189–F213.

Caliendo, M. (2006). *Microeconometric evaluation of labour market policies.* Springer Verlag.

Carneiro, P., Hansen, K., and Heckman, J. (2001). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Economic Policy Review*, **8**, 273–301.

Carneiro, P., Hansen, K., and Heckman, J. (2003). Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College. *NBER working paper.*

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, **71**, 805–819.

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, **88**, 1007.

Cook, T. (2008). Waiting for life to arrive: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, **142**, 636–654.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220.

Crump, R., Hotz, V., Imbens, G., and Mitnik, O. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, **90**, 389–405.

Cunha, F., Heckman, J., and Navarro, S. (2005). Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers*, **57**, 191.

Cunha, F., Heckman, J., and Navarro, S. (2006). Counterfactual analysis of inequality and social mobility. *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, pages 290–348.

Cunha, F., Heckman, J., and Navarro, S. (2007). The identification and economic content of ordered choice models with stochastic cutoffs. *International Economic Review*, **48**, 1273–1309.

Dawid, A. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–189.

DeFillippi, R.J., A. M. (1996). *Boundaryless contexts and careers: A competency based perspective. In Arthur, M.B. and Rousseau, D.M. (eds) The boundaryless career: A new employment principle for a new organizational era.* New York: Oxford University Press.

Dempster, A., Laird, N., Rubin, D., *et al.* (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1–38.

Donald, S. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, **89**, 221–233.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, **68**, 589.

Fabrizi, A. (2009). Le public knowledge partnerships nell'ambito delle politiche per la ricerca e lo sviluppo tecnologico dell'Unione Europea.

Fisher, R. (1925). *The design of Experiments*. 1st ed, Oliver and Boyd, London.

Freedman, D. (1999). From association to causation: some remarks on the history of statistics. *Statistical Science*, **14**, 243–258.

Gagliarducci, S. (2005). The dynamics of repeated temporary jobs. *Labour Economics*, **12**, 429–448.

Gaure, S. Zhang, T. and Røed, K. (2007). Time and causality: A Monte Carlo assessment of the timing-of-events approach. *Journal of Econometrics*, **141**, 1159–1195.

Gill, R. and Robins, J. (2001). Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, **29**, 1785–1811.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, **37**, 424–438.

Gritz, R. (1993). The impact of training on the frequency and duration of employment. *Journal of Econometrics*, **57**, 21–51.

Güell, M. and Petrongolo, B. (2007). How binding are legal limits? Transitions from temporary to permanent work in Spain. *Labour Economics*, **14**, 153–183.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, **11**, 1–12.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69**, 201–209.

Hansen, L. and Sargent, T. (1980). Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control*, **2**, 7–46.

Heckman, J. (1981). 3 Heterogeneity and State Dependence. *Studies in labor markets*, page 91.

Heckman, J. (1992). Randomization and Social Policy Evaluation. S. 201-230 in Charles F. Manski und Irwin Garfinkel (Hrsg.), Evaluating Welfare and Training Programs.

Heckman, J. (2005). The scientific model of causality. *Sociological methodology*, **35**, 1.

Heckman, J. (2008). Econometric causality. *International Statistical Review*, **76**, 1–27.

Heckman, J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, **48**, 356–398.

Heckman, J. and Honore, B. (1990). The empirical content of the Roy model. *Econometrica: Journal of the Econometric Society*, **58**, 1121–1149.

Heckman, J. and Hotz, V. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, **84**, 862–874.

Heckman, J. and Navarro, S. (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, **136**, 341–396.

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*, **86**, 30–57.

Heckman, J. and Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions:: An overview. *Journal of Econometrics*, **30**, 239–267.

Heckman, J. and Singer, B. (1984). Economic duration analysis. *Journal of Econometrics*, **24**, 63–132.

Heckman, J. and Singer, B. (1986). Econometric analysis of longitudinal data. *Handbook of econometrics*, **3**, 1689–1763.

Heckman, J. and Smith, J. (1998). Evaluating the welfare state. *NBER working paper*.

Heckman, J. and Vytlacil, E. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4730.

Heckman, J. and Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, **91**, 107–111.

Heckman, J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, **73**, 669–738.

Heckman, J. and Vytlacil, E. (2007a). Econometric evaluation of social programs. *Handbook of Econometrics*, **6B**.

Heckman, J. and Vytlacil, E. (2007b). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, **6B**, 4779–4874.

Heckman, J. and Vytlacil, E. (2007c). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics*, **6B**, 4875–5143.

Heckman, J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, **64**, 487–535.

Heckman, J., Lochner, L., Taber, C., Foundation, A., Str, P., and Hall, H. (1998a). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *NBER working paper*.

Heckman, J., Lochner, L., and Taber, C. (1998b). General-equilibrium treatment effects: A study of tuition policy. *American Economic Review*, **88**, 381–386.

Heckman, J., Lochner, L., and Taber, C. (1998c). Tax policy and human-capital formation. *American Economic Review*, **88**, 293–297.

Heckman, J., LaLonde, R., and Smith, J. (1999). The economics and econometrics of active labor market programs. *Handbook of labor economics*, **3**, 1865–2097.

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

Holland, P. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological methodology*, **18**, 449–484.

Hsiao, C. (2003). *Analysis of panel data*. Cambridge Univ Pr.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, **62**, 467–475.

Imbens, G. and Lemieux, T. (2008). Special issue editors' introduction: The regression discontinuity design–Theory and applications. *Journal of Econometrics*, **142**, 611–614.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, **47**, 5–86.

Kamakura, W. and Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, **26**, 379–390.

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

Lancaster, T. (1990). The econometric analysis of transition data, Econometric Society Monograph No. 17.

Lechner, M. (2004). Sequential matching estimation of dynamic causal models. *Institute for the Study of Labor (IZA); University of St. Gallen-Research Institute for Empirical Economics and Economic Policy; Centre for Economic Policy Research (CEPR)*.

Lechner, M. (2006). The relation of different concepts of causality in econometrics. *Institute for the Study of Labor (IZA); University of St. Gallen-Research Institute for Empirical Economics and Economic Policy; Centre for Economic Policy Research (CEPR)*.

Lechner, M. and Miquel, R. (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics*, pages 1–27.

Lee, D. (2001). The electoral advantage to incumbency and voters' valuation of politicians' experience: a regression discontinuity analysis of elections to the US House.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**, 226–233.

Lucidi, F. and Raitano, M. (2009). Molto flessibili, poco sicuri: Lavoro atipico e disuguaglianze nel mercato del lavoro italiano. *Economia e Lavoro*, **2**, 99–115.

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

Manski, C. (1995). *Identification problems in the social sciences*. Harvard Univ Pr.

Manski, C. (2003). *Partial identification of probability distributions*. Springer Verlag.

Manski, C. (2005). *Social choice with partial knowledge of treatment response*. Princeton Univ Pr.

Manski, C. (2007). *Identification for prediction and decision*. Harvard Univ Pr.

Marschak, J. (1953). Economic measurements for policy and prediction. *Studies in econometric method*, pages 1–26.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**, 109–142.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley-Interscience.

Mezzanzanica, M. (2008). La mobilità del mercato del lavoro in provincia di Milano. *Impresa e Stato: rivista della Camera di commercio di Milano*.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9 (1923). *Translated in Statistical Science*, **5**, 465–480.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge Univ Pr.

Pearl, J. (2010). An introduction to Causal Inference. *The International Journal of Biostatistics*, **6**.

Quandt, R. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53**, 873–880.

Quandt, R. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, **67**, 306–310.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512.

Robins, J. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, **113**, 159.

Robins, J. (1997). Causal inference from complex longitudinal data. *Latent variable modeling and applications to causality*, **120**, 69–117.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Rosenbaum, P. (1987). The role of a second control group in an observational study. *Statistical Science*, **2**, 292–306.

Rosenbaum, P. (1995). Observational studies. *NY: Springer-Verlag*.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41.

Rosenzweig, M. and Wolpin, K. (2000). "Natural experiments" in economics. *Journal of Economic Literature*, **38**, 827–874.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, **3**, 135–146.

Rubin, D. (1973a). Matching to remove bias in observational studies. *Biometrics*, **29**, 159–183.

Rubin, D. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

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

Rubin, D. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational and Behavioral statistics*, **2**, 1.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, **6**, 34–58.

Rubin, D. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, **81**, 961–962.

Rubin, D. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, **25**, 279–292.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

Shadish, W., Cook, T., and Campbell, D. (2002). Experimental and quasi-experimental designs for generalized causal inference.

Sims, C. (1972). Money, income, and causality. *The American Economic Review*, **62**, 540–552.

Ten Have, T., Joffe, M., and Cary, M. (2003). Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine*, **22**, 1255–1283.

Thurstone, L. (1927). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, **21**, 384–400.

Tukey, J. (1986). Comments on alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. *Drawing Inferences from Self-Selected Samples*, pages 108–110.

Van den Berg, G. (2001). Duration models: specification, identification and multiple durations. *Handbook of econometrics*, **5**, 3381–3460.

Van den Berg, G., Van der Klaauw, B., and Van Ours, J. (2004). Punitive sanctions and the transition rate from welfare to work. *Journal of Labor Economics*, **22**.

Van der Klaauw, W. (2008). Regression–discontinuity analysis: a survey of recent developments in economics. *Labour*, **22**, 219–245.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482.

Wilks, S. (1935). The likelihood test of independence in contingency tables. *The Annals of Mathematical Statistics*, **6**, 190–196.

Yule, G. (1899). An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades (Part I). *Journal of the Royal Statistical Society*, **62**, 249–295.

Zijl, M., Van Den Berg, G., and Heyma, A. (2004). Stepping stones for the unemployed: the effect of temporary jobs on the duration until regular work. *IZA Discussion Papers*, **1241**.