Università degli Studi di Firenze

Dipartimento di Statistica "G. Parenti"

Dottorato di Ricerca in Statistica Applicata

XX ciclo – SECS-S/01

Causal inference for observational studies extended to a multilevel setting. The impact of fertility on poverty in Vietnam

Bruno Arpino

Tutor: Prof. Fabrizia Mealli

Co-tutor: Prof. Arnstein Aassve, Prof. Letizia Mencarini

Coordinatore: Prof. Guido Ferrari

Aknowledgments

A journey is easier when you travel together. Interdependence is certainly more valuable than independence. Apart from the statistical complications! This thesis is the result of a work whereby I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

Foremost, I would like to thank my supervisors Professor Fabrizia Mealli, Professor Arnstein Aassve and Professor Letizia Mencarini. I am very appreciative of their generosity with advice, motivations and references, to name a few of their contributions. They brought unique perspectives to my research, enriching it greatly. Without their support, this work would not have been the same.

I especially want to thank my first supervisor Professor Fabrizia Mealli for her guidance throughout all my thesis work. I benefited during the PhD period by her immense knowledge in statistics and, especially, causal inference, her ability to formalise problems, her common-sense. I am very grateful to Professor Letizia Mencarini for her insightful suggestions from a demographic perspective. I deserve special thanks to Professor Arnstein Aassve for his patience, motivation, enthusiasm for my research. A part of my work was carried out when I was visiting the Institute for Social and Economic Research (ISER) of the University of Essex. I spent at ISER two very stimulating research periods under the supervision of Professor Aassve, profiting by his experience in econometrics, demographic and poverty research and ability in data analysis.

I also owe a great deal of gratitude to a number of people who read parts of this work, giving me valuable comments: Dr. Francesca Francavilla, Professor

Leonardo Grilli, Professor Stefano Mazzucco, Sonca Nguyen, Professor Stephen Pudney, Professor Carla Rampichini.

Preliminary papers based on parts of this work was presented at some conferences: 56th session of the International Statistical Institute (Lisbon, 22-29 August); 48ª Riunione Scientifica Annuale della Società Italiana degli Economisti (Turin, 26-27 October); 2nd International Workshop on poverty and social exclusion: dynamics and multidimensional issues (Barcelona, 24 November). I am grateful to participants at these conferences for critics and comments that improved the work.

Finally, I would like to thanks all the staff of the Statistical Department of the University of Florence and my colleagues, who gave me the feeling of being at home at work.

Contents

Preface

The relationship between fertility and poverty is a topic studied by economists and demographers for a long time. The very first researches on the linkage between population and poverty adopted a macro perspective, that is, the topis was studied at the national or state level. In this context, the neoclassic economic theories argue that population growth has a negative impact on economy due to, mainly, decreasing marginal returns of work and to existing obstacles to capital accumulation. In the last decades, the micro approach, which usually takes the household as unit of analysis, has been remarkably developed. In this context, the crucial aspect of the interaction between the quantity and quality of children was introduced (Becker and Lewis, 1973; Barro and Becker, 1989; Becker et al, 1999). Children are considered as an essential part of the household's work force as they generate income, as well as providing insurance against old age. This is especially true for male children and for households living in rural underdeveloped regions (Admassie, 2002).

Existing micro-level researches on the relationship between poverty and fertility in Less Developed Countries (LDC) are mainly based on cross-sectional data. The results vary considerably (Schoumaker and Tabutin, 1999). However, the most common relationship between poverty and fertility, in contemporary LDC, is positive. These results underlie the presumption of a positive causal relation between poverty and fertility at the household level. Whereas there is a clear positive *association* between fertility and poverty, it is not equally clear to what extent fertility actually *leads* to a worsened economic situation. This is of course a very different question, since we are in this case interested in the *causal* effect of fertility on poverty, which ultimately is what we would need in order to give sound policy advice. Policy-makers are naturally interested in causality.

Good public policy decisions require reliable information about the causal relationships among variables. Policy-makers must understand the way the world works and the likely effects of manipulating the variables that are under their control. If, for example, having more children causes poverty, policy makers could, adequately, plan some actions to impact on fertility, directly or indirectly. Alternatively, if the only policy goal is to contrast poverty conditions without any wish to determine fertility behaviours, it could be simply decided to compensate the higher costs supported by households with a lot of children through state benefits.

In order to draw proper causal conclusions about the effect produced by a social phenomenon on another we need to use appropriate statistical methodologies and data. In the literature, there are few studies that approached the fertility-poverty relationship from a causal perspective using adequate methodologies. Moreover, only recently panel data on LDC are made available due to the implementation of Living Standards Measurement Surveys (LSMS) conducted in a number of countries with technical assistance from the World Bank.

Likewise all LSMS, the Vietnamese surveys (VLSMS) include rich information on variables that are important determinants for the household's standard of living and fertility behaviour. For example, it collects data on education, employment, fertility and marital histories, together with detailed information on household income and consumption expenditure. A very interesting feature of the VLSMS is that it also provides, for the rural areas, detailed community information from a separate questionnaire.

The longitudinal dimension of the data available is crucially important to be allowed to draw robust causal inference about the effect of interest. In fact, only longitudinal data allow us to keep into account the dynamic nature of fertility and poverty processes. By using data on two time points we properly can implement a pre-post treatment analysis which is vital for our study of causal inference.

On the other side, as already mentioned, we need to use adequate statistical methodologies for causal inference. Despite other situations, the study

of the relationship between fertility and poverty cannot rely on experiments. On the contrary, we can use survey data, such those coming from the VLSMS, and adopt a quasi-experimental approach for our observational study.

The approach to causal inference we adopt is the potential outcomes framework, pioneered by Neyman (1923) and Fisher (1925) and extended by Rubin (1974, 1978) to observational studies. Recently, the approach has been adopted by many in both statistics and econometrics (e.g. Rosenmbaum and Rubin, 1983a; Heckman, 1992 and 1997a; Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Heckman, Ichimura and Todd, 1997). This literature formalises notions of cause and effect and is based on the counterfactual idea. Counterfactual refers to what would have happened if, contrary to fact, the exposure had been something other than what it actually was (Greenland and Brumback, 2002). As a means to show the idea, suppose we have a population of individual units under study (in our case households) indexed by $i = 1, 2, \ldots N$, a treatment indicator D, that assume the value 1 for treated units and 0 for untreated or the controls and an outcome variable, here indicated by Y. Each unit, i, has two potential outcomes depending on its assignment to the treatment levels: Y_{i1} if D_i=1 and Y_{i0} if D_i=0. The fact that potential outcomes for each unit depends only on the treatment received by that unit corresponds to the "*no interference among units*" assumption of Cox (1958), which Rubin (1980) refers to and extends as the Stable Unit Treatment Value Assumption (SUTVA).

As with many of the other assumptions to be discussed, it is important to note that SUTVA is not directly informed by the data. In other words, it is an un-testable assumption that stems from the scientist's assessment or knowledge.

Following Rubin (1978), each comparison (obtained through means, ratios and so on) among potential outcomes defines a causal effect of potential interest. It is obvious that the two potential outcomes are not observables for the same unit - a feature referred to by Holland (1986) as the "*fundamental problem of causal inference*". In so far, causal inference can be seen as a missing data problem.

Generally, we cannot draw valid causal conclusions without considering what makes some units receive a treatment whereas others do not. This is

referred to as the *assignment mechanism*, and there is a critical distinction between randomized and observational studies. The key difference is that in randomized settings, the analyst can control assignment to treatment and the probabilities of being assigned to treatment are known. In observational studies, as ours, these conditions are unlikely to hold and the researcher can only estimate probabilities of assignment to treatment on the basis of the data available. By adopting the potential outcomes framework and by using panel data from the VLSMS, we review the fertility-poverty relationship from a causal perspective.

Another key perspective that we take in this work is a multilevel one. The multilevel approach is motivated by the consideration that the place where households reside has important consequences both for their poverty and fertility conditions. In particular, households can be considered as clustered in communities. This implies a two-level data structure, with households at the first level and communities at the second.

Keeping explicitly into account this multilevel dimension in the study of the causal effect of fertility of poverty is central, both for statistical and for substantive research reasons. The fact that community characteristics (infrastructure, remoteness, culture and so on) influence both phenomena requires to control also for them in the statistical analyses. Otherwise, we might capture associations that are not causal. Moreover, the multilevel dimension implies specific challenges for causal inference. Apart from the statistical motivations, the multilevel structure of the data brings some interesting research question. For example, it is of interest to understand if the effect of fertility on poverty changes by community.

Outline of the thesis

The outline of the thesis follows the logical development of the previous discussion. We start (chapter 1) by introducing the background of our empirical analyses. The principal concepts and measures of poverty are briefly discussed. We use a very standard and consolidated approach for poverty measurement in

LDC based on households' consumption expenditure as a measure of welfare. Then, we discuss the main determinants of poverty and fertility, as they are analyzed in the theoretical and empirical literature on the topic. Understanding the common determinants of the two phenomena is vital for our study of causal inference. Failing to control for these variables hamper any detection of causal effects. We, briefly, discuss the previous works studying the fertility-poverty relationship, stressing that they often fail to use adequate data and methodologies. Then, we introduce the data we use in our applications. As already mentioned, they come from the VLSMS, which is a panel consisting of two waves covering the 1990s, a period of strong economic growth for Vietnam. We clarify that the process of the improvement in the economic and social conditions is imputed, by many observers, to the "Doi Moi" (renovation) policy implemented from 1980s by the Vietnamese government. Finally, chapter 1 concludes by explaining the motivations underlying the need for adopting a causal and a multilevel perspective in our work.

We continue (chapter 2) presenting the potential outcomes framework, under which our causal inference is made. We show the basic concepts and definitions of causal effects and assignment mechanism. We categorise different situations for causal inference, distinguishing among *randomised* and *observational studies*. Among observational studies we distinguish two situations referred to as *regular* and *irregular* assignment mechanisms. The first concerns studies where the analyst can reasonably assume that characteristics driving selection into treatment are all observed. Among irregular assignment mechanism the most important case is represented by the *latent regular assignment* where selection also depends on unobserved characteristics. Randomized experiments with non compliance, and by extension, instrumental variables estimation, belong to this setting. We review several methods for causal inference that we can use in the manifold outlined situations, stressing the differences among them in terms of assumptions and data requirement. This review includes, in particular, recent methodology for using instrumental variables with covariates avoiding traditional methods, which often rely on strong assumptions.

In chapter 3, we present the general motivations for using a multilevel approach to the study of social phenomena. The traditional multilevel linear model is reviewed, focussing on the second level endogeneity problem. By using the potential outcomes framework, we originally re-analyse multilevel linear models, stressing the modifications needed to the standard framework when causal inference is made in a multilevel setting. Some pitfalls of these models in recovering causal effects are emphasised. Finally, we discuss the statistical and substantive motivations to keep explicitly into account the multilevel dimension. Three vital topics for causal inference in a multilevel setting are, in particular, explored.

The first one refers to the cluster-heterogeneity in the treatment effect. Actually, this issue is not a statistical one but it is driven from a research question about the heterogeneity of the treatment effect. Moreover, it is not specific to multilevel settings. In all studies of causal inference we could be interested in the treatment effect heterogeneity. However, in a multilevel setting it could be of specific interest to learn if and why some heterogeneity in the treatment effect is driven by the characteristics of the cluster to which units belong.

The second issue concerns the fact that, in a multilevel setting, not only the outcome model, but also the selection process can have a multilevel structure. This is the case when the probability of being treated changes substantially by cluster, and the effect of some covariates on this probability varies by cluster. The first aspect requires the inclusion of a random intercept in the model of the propensity score, while the second one asks for the inclusion of random slopes. In other words, the statistical implication of the multilevel structure of the selection process is that including only observed covariates in the model for the propensity score, and hence balancing only for them, might not be sufficient. Some unobserved cluster level characteristics could be related to both the treatment and the outcome, generating bias in the estimation of causal effects. In this context, we propose a two-stage strategy to allow balancing observed covariates, defined both at the first and at the second level, as well as

empirical bayes predictions of the random effects, which capture unobserved effects at the cluster level.

Finally, we consider the potential invalidity of the SUTVA in a multilevel setting. In general, this assumption is problematic when sharing and competition for resources generate interference among units (at least) belonging to the same cluster. Inference without SUTVA is complicated since potential outcomes for each unit depend also on the treatment received by the other units. In a multilevel setting, this problem is traditionally overcome by redefining the unit of analysis at the minimum aggregate level for which the assumption is tenable. However, the consequence is that the analysis should be conducted at an aggregate level and we cannot refer our results to the individual level. Otherwise, we could commit an ecological fallacy error. Since in our application, as it is often the case in multilevel analyses, we are interested in drawing inference at the unit level we need a weaker version of SUTVA that allow us to conduct the study still at the first level. We discuss a weaker version of the SUTVA, which amounts to assume that there is no interference among units belonging to different clusters, while the within-cluster interference is fully captured by the level of the proportion of treated (high versus low).

In the following three chapters we present the empirical part of our work. Chapter 4 contains a multilevel analysis of poverty exit determinants in Vietnam. This application is appealing *per se*. However, in the economy of the present work, it is interesting to explore the between-communities variability in the change of household living standards. If there is substantive between-communities variability, even after controlling for observed covariates, then a multilevel analysis is justified. We would like to learn about the key determinants of the transition from poverty to non-poverty, keeping into account the clusterisation of household in communities. An attractive development of the analysis is the proposal of using empirical bayes prediction of the random effect to help policy makers to better calibrate their decisions.

In chapter 5, we apply some of the methods discussed in the chapter 2 to the problem of estimating the causal effect of childbearing on changes in households' consumption expenditures. The issue is that childbearing events

cannot be considered as an exogenous measure of fertility, especially when the outcome relates to economic wellbeing – in our case measured in terms of consumption expenditures. We, first, contrast some methods based on the unconfoundedness assumption: regression and propensity score matching. We then assess the potential effect from omitting relevant but unobserved variables without actually implementing an Instrumental Variable (IV) approach, through an extended sensitivity analysis. This is a very useful tool, in the sense that valid and relevant instruments are often hard to come by. However, in our application we also explore the use of the IV method using two different instruments. The first is a well-used instrument that relates to couples' preference for sons. The second instrument is related to the contraceptives availability in the community where household reside. Since this second instrument cannot be thought as random we need to control for covariates. We explore the use of a recent approach suggested by Frölich (2007), which overcomes many of the stringent assumptions typical of the traditional IV methods with covariates. We use this application as a means to illustrate the existing difference among methods based on different assumptions. In particular, we contrast regressions and PSM versus IV.

In chapter 6, we re-analyse the estimation of the causal effect of fertility on poverty, treated in the previous chapter, with the goal of keeping into account the multilevel dimension of the problem. First, we analyse the effect of fertility on poverty using multilevel models. Then, we adopt a different strategy consisting in a combination of multilevel models for the estimation of the propensity score and matching methods for the estimation of causal effects. We compare different strategies for the specification of the propensity score, which are evaluated with reference to the balance they allow us to achieve in observed covariates and in the prediction of random effects included in multilevel models. Finally, we explore the complication due to the potential violation of the SUTVA in a multilevel setting. We compare results obtained under the standard version of this assumption with those we get under a weaker version. We conclude this work with some overall considerations on the key empirical and methodological results we found.

Chapter 1
Analysing the causal effect of fertility on poverty in Vietnam

Introduction

A common observation in developing countries is that large households with many children tend to be poorer. Whereas there is a clear positive association between fertility and poverty, it is not equally clear to what extent fertility actually *leads* to a worsened economic situation. This is of course a very different question, since we are in this case interested in the causal effect of fertility on poverty, which ultimately is what we would need in order to give sound policy advice.

The main thesis goal is to use advanced statistical techniques in order to estimate the causal effect of fertility on poverty. In this chapter we introduce the background and discuss the perspectives we used in this work.

The chapter is organized as follows. Section 1.1 briefly discusses the principal concepts of poverty and the view we adopt in this work. Sections 1.2 and 1.3 examine, respectively, the theoretical determinants of poverty and fertility. This discussion is important to understand which are the most important variables potentially correlated with both phenomena. Failing to control for these variables hamper any detection of causal effects. Section 1.4 offers a short review of the existing literature about the relationship between poverty and fertility. Section 1.5 presents the data we used in the thesis and the Vietnamese context. Section 1.6 motivates the perspectives we adopted to develop this work. Section 1.7 concludes.

1.1 Concepts and measures of poverty

According to the World Bank (2000), "poverty is pronounced deprivation in well-being." This of course begs the question of what is meant by well-being. The conventional view links well-being primarily to command over commodities, so the poor are those who do not have enough income or consumption to put them above some adequate minimum threshold. This view sees poverty largely in monetary terms. Poverty may also be tied to a specific type of consumption; thus someone might be house poor or food poor or health poor. These dimensions of poverty can often be measured directly, for instance by measuring malnutrition or literacy.

The broadest approach to well-being (and poverty) was pioneered by Sen (1987) and focuses on the "capability" of the individual to function in society. The poor lack key capabilities, and may have inadequate income or education, or be in poor health, or feel powerless, or lack political freedoms. Viewed in this way, poverty is a multi-dimensional phenomenon, and less amenable to simple solutions. So, for instance, while higher average incomes will certainly help reduce poverty, these may need to be accompanied by measures to empower the poor, or insure them against risks, or to address specific weaknesses (such as inadequate availability of schools or a corrupt health service).

Recognizing the multidimensional nature of poverty is important when we want to analyse living standard conditions or dynamics in a given area. The definition of poverty rates based only on monetary (consumption or income) data can be misleading or insufficient to describe the complexity of the reality of poor's conditions. This argument is valid in developed as well as Less Developed Countries (LDC), like Vietnam. However, in LDCs the monetary dimension of well-being assumes a higher relative weight. Moreover, multidimensional measures of poverty are more difficult, requiring more detailed information and sophisticated methods. In our work we adopt the first concept of poverty which seems quite adequate to study the situation of the rural Vietnam and allow us to use a standard and consolidated measure of poverty.

The first step in measuring poverty is defining an indicator of welfare such as income or consumption. Information on welfare is derived from survey data. The World Bank-inspired Living Standards Measurement Surveys (LSMS) feature multi-topic questionnaires and strict quality control. The flexible LSMS template is widely used. We used in this work the Vietnamese LSMS we present in section 1.5.

Income is generally used as a measure of welfare in developed countries, but tends to be seriously understated in less-developed countries (Coudouel et al., 2002; Deaton and Zaidi, 2002). Consumption is less understated and comes closer to measuring "permanent income." However, it requires detailed information on consumption behaviours, their expenditure pattern and the evaluation of durable goods (by assessing the implicit rental cost) and housing (by estimating what it would have cost to rent).

While consumption per capita is the most commonly-used measure of welfare, some analysts use consumption per adult equivalent, in order to capture differences in need by age, and economies of scale in consumption. The standard solution is to impose an assumption on inter-household resources allocation, and adjustments can be done by applying an equivalence scale that is consistent with the assumption made – producing a measure of expenditure per equivalent adult.

As a measure of household's living standard, we use the household's consumption expenditures using the expenditure variables calculated by the World Bank procedure which is readily available with the Vietnam LSMS (VLSMS) survey. We apply a simple equivalence scale similar to the one White and Masset (2002) estimated on VLSMS, giving to each child aged 0-14 in the household a weight of 0.65 relative to adults. This means that the mean poverty rate for the two waves will be different from the official ones, given that the latter is based on per capita expenditure, which in effect implies an equivalence scale assigning equal weights to all household members.

When using consumption or income data we have also to bear in mind that households face different prices according to the place they live. Moreover, since we use two waves surveyed in different time points, we have to consider also the variability of prices over time. In order to keep into account both

aspects, we opportunely deflate households' consumption expenditure by means of price indexes available with the VLSMS.

1.2 The determinants of poverty

Poverty and poverty reduction are currently the central concerns of development discourse and policy-makers agenda. Although the construction of poverty profiles is useful because it allows us to know whether poverty is increasing or decreasing, as well as the changes in the composition of the population in poverty, poverty profiles do not throw much light about the causes of poverty.

In fact, a country poverty profile simply sets out the major facts on poverty (and typically, inequality), and then examines the pattern of poverty, to see how it varies by geography (by region, urban/rural, mountain/plain, etc.), by community characteristics (e.g. in communities with and without a school, etc.), and by household characteristics (e.g. by education of household head, by household size). They only provide a description of poverty according to several economic, demographic or social characteristics, but do not go in depth as to look for the underlying causes of differences in poverty rates across population groups and/or across time. Understanding the determinants of poverty is crucially important for policy making. We discuss this issue, specifically the need for a causal perspective, in section 1.6.

Few questions have generates much discussion across time as that of the causes of poverty. The sources and origin of poverty have been debated for centuries. In fact, almost as long as there has been poverty in the world, there have been attempts to explain it with the ultimate goal of alleviate poor conditions. As the historian Hartwell (1986) notes, ''The causes of poverty, its relief and cure, have been a matter of serious concern to theologians, statesmen, civil servants, intellectuals, tax payers and humanitarians since the middle-age''.

In general, empirical studies on the poverty determinants have used different methodologies, including ordinary least square regression where the dependent variable is continuous (e.g. logarithm of consumption expenditures),

logistic regression where the dependent variable is binary (poor or non poor), and quantile regressions (Garza-Rodríguez, 2004). The adequacy of regression approach to causal inference will be discussed in chapter 2.

However, it is important to recognize that these methodologies can only corroborate or contradict, empirically, a theoretically defined relationship among poverty and a group of variables, assumed to be its determinants. In this section we briefly review which are the most important theoretically determinants of poverty as they are stated by economic, sociologic and other theoretical studies. The empirical literature, obviously, it is important since it has served to confirm the theoretical beliefs.

The determinants of poverty can be distinguished in three groups: individual, household and aggregate. Individual and household determinants can be organized into three groups: demographic, economic and social characteristics. Important individual characteristics are without doubt represented by age, sex and marital status. Also the demographic composition of the household is a key factor: number of children, of elderly members, the dependency ratio (calculated as the ratio of the number of family members not in the labour force to those in the labour force in the household). Also the gender of the household head can significantly influences household poverty. In many LDCs females are discriminated in the labour market and generally in the society. This fact impacts directly on females' living standards and on female-headed households' conditions.

As economic characteristics are concerned, apart from income or consumption, which are typically used to define whether a household is poor, there are a number of other economic aspects important for individual and household poverty. First of all, employment status including the type and sector of activity is an important aspect to be considered in a poverty analysis. Another important feature refers to the assets owed by household members, including the house, land, livestock, and agricultural machinery. However, these assets, at least some of them, could also be included in a broader measure of well-being.

Aside from the demographic and economic indicators, several social characteristics are important for living standards. The most critical are measures

of health and education. Nutritional and disease status, are examples of indicators normally used to characterize health in analyzing living standards. As education is concerned, literacy and schooling are important indicators of the quality of life in their own right, as well as being key determinants of poor people's ability to take advantage of income-earning opportunities. Race and religion are other factors to be considered in those countries where some form of discrimination against minority is in action. This is the case of Vietnam where some of the ethnic groups different from the Kinh are segregated.

In addition to the individual and household characteristics, the literature has been placing an increasing focus on the role of the features of the place where households reside (e.g. Van de Walle, 1996; Glewwe et al, 2002; Ali and Pernia, 2003; Mukherjee and Benson, 2003; Justino and Litchfield, 2004).

There are a lot of geographical and institutional features that impact considerably on poverty. At the regional level, there are numerous characteristics that might influence peoples' living standards. The relationship of these characteristics with poverty is country-specific. In general, however, poverty is high in areas characterized by geographical isolation, a low resource base, low rainfall, and other inhospitable climatic conditions. Vietnam is poor in part because it is regularly hit by typhoons, which destroy a significant part of the accumulated stock of agricultural capital. In many parts of the world the remoteness of rural areas (which lower the price farmers get for their goods and raise the price they pay for purchases, due to high transport costs) is responsible for generating food insecurity among the poor. Inadequate public services, weak communications and infrastructure, as well as underdeveloped markets are dominant features of life in many rural parts of the world, and clearly contribute to poverty.

Other important regional and national characteristics that affect poverty include good governance, economic, political and market stability, mass participation, global and regional security, intellectual expression and a fair, functional, and effective judiciary. Regional-level market reforms can boost growth and help poor people.

As with regional characteristics, there are a variety of community-level characteristics that may influence poverty. Among them, infrastructures are a major determinant of peoples' economic conditions. Indicators of infrastructure development that have often been used in econometric exercises include proximity to paved roads, whether or not the community has electricity, proximity to large markets, availability of schools and medical clinics in the area, and distance to local administrative centres. Other indicators of community level characteristics include average human resource development, access to employment, social mobility and representation, and land distribution. The VLSMS include, as we will see in section 1.5, for the rural part of Vietnam essential data on community characteristics allowing to include an important part of the information into the analysis.

Recent research has also stressed the importance of social networks and institutions, and "social capital" (which includes, for instance, the level of mutual trust in the community). Social institutions refer to the kinship systems, local organizations, and networks of the poor and can be thought of as different dimensions of social capital. Research on the roles of different types of social networks in poor communities confirms their importance.

In general, we cannot deny the role that the geographical environment (physical, as well as social) plays in the formation of all sorts of human behaviour including economic behaviours in a broad sense (Skinner, 1965). This is the case also for fertility behaviour. As we will argue in the next section, also for fertility the geographical dimension plays a crucial role.

If all the mentioned factors can be viewed as causes of poverty is not always clear. There are reasons to raise questions about the direction of causality for a lot of these variables, that is, in several cases authors have raised questions such as whether poverty influences the level of the variable in question, or whether the variable in question is not an intervening variable for some more fundamental determinant of poverty. In this work we are specifically interested in the relationship between poverty and fertility. If is fertility to cause poverty or vice-versa is matter of debate, as we will see in section 1.4.

1.3 The determinants of fertility

Fertility makes possible the continuity of populations, societies and cultures. It is central, ineluctable and remarkably complex. As much as any human activity, reproduction involve biology and culture, individual and communal. While numbers of births over time have significant consequences for population size and structure, its analysis cannot be only demographic. Its study has been approached, in fact, by different branch of social science: economics, anthropology, sociology.

In any discussion of fertility, it is useful to begin with the recognition of the role of intermediate variables, through which any social factors influencing the level of fertility must operate. The intermediate fertility variables, also called the "proximate determinants" of fertility, offer a way of attributing the variation in fertility to specific mechanisms. A variety of explanatory variables, including biological variables, infant and child mortality, the role of women, education, and access to resources, have been shown to directly or indirectly explain observed variations in fertility. Davis and Blake (1956) provide a taxonomy of mutually exclusive intermediate variables suggesting that there are 3 categories of variables that are necessary for successful reproduction: variables which define the probability of sexual intercourse, such as age of entry into sexual unions; variables which define the probability of a conception resulting from sexual intercourse, such as contraceptive use and fecundity; and variables which define the probability of a conception resulting in a live birth, such as spontaneous or induced abortion.

According to Easterline (1975), in the microeconomic theory the determinants of fertility are seen as working through one or more of the following processes: 1) the demand for children, that is the number of surviving children parents would want if fertility regulation were costless; 2) the potential output of children, that is the number of surviving children parents would have if they did not deliberately limit fertility; and 3) the costs of fertility regulation, including both subjective (psychic) costs and objective costs, the time and money required to learn about and use specific techniques.

Economists in their research about the determinants of fertility have typically emphasized the following variables: age; mortality; budget constraints; the status of women; education; direct costs and benefits of children. Another category of variables relates to the environment within which family decision making is undertaken. There are two basic kinds of variables which we can include under this heading: variables relating to the political, social, and economic status of the community as a whole; and variables relating to specific policies and programs which are likely to have a direct or indirect influence on population or one of its components.

As happened for poverty, in the literature, the role of the characteristics of the place where households reside have been received an increased attention (e.g. Entwisle et al, 1989; Hirschman and Guest, 1990; Josipovic, 2003). This fact derives from the recognition that the human fertility is a socially modified biological process. This social modification of fertility is the consequence of numerous groups of factors that issue directly from society or are its product.

From the viewpoint of geographical factors of fertility, where a person lives is significant since to what extent she will realize her physiological fecundity also depends on the place (that is, on the relief, the transportation infrastructure, the distance from central settlements, accessibility to various facilities, the economic activities, the quality of the environment and living conditions, the level of urbanization, and similar factors).

In the field of fertility behaviour, geographical differences can occur due to the specific regional-geographical structure or due to the different strength of individual factors. The strength of an individual factor is linked to the place where it occurs. Thus, each indirect fertility factor has its spatial or regional component that reflects its differential strength or spatial or regional differentiation.

Summarizing, the determinants of individual fertility, likewise those of poverty as we have seen in the previous section, can be distinguished in individual, household and geographical/institutional factors. This motivates, as we better discuss in section 1.6 and chapter 3, the need for a multilevel perspective to the study of these phenomena.

1.4 The literature about the relationship between fertility and poverty

The relationship between fertility and poverty is a topic studied by economists and demographers for a long time. As Livi-Bacci (1994) observes, there are mainly two ways of dealing with the relation between the two phenomena. The first approach analyzes this relation at a macro level, that is, at the level of country or region. In this framework, the main empirical finding is related to the observation of a positive relation between the rate of population growth and the incidence of poverty. The second approach analyzes the relation at the household or individual level. This is the approach that we employed in our research for the Vietnam context taking the household as the statistical unit of analysis.

The very first researches on the linkage between population and poverty adopted a macro perspective. We cannot fail to remember the celebrated work of Malthus (1798), *An essay on the principle of population,* which can be considered as the first organic work on the impact that population growth has on economy. On the basis of some well-known assumptions, Malthus explained the ineluctable tension existing among available resources and demographic trends. The Malthusian model continues, in part, to live in the neoclassic economic theories. These argue that population growth has a negative impact on economy due to decreasing marginal returns of work and to obstacles existing to capital accumulation.

In the last decades the micro approach has been remarkably developed. In this context we have to remember the fundamental work of Becker and colleagues (Becker and Lewis, 1973; Barro and Becker, 1989; Becker et al, 1999). They introduced in the economics of family the crucial aspect of the interaction between the quantity and quality of children. In these theoretical works, the household utility function include as argument both the quantity and quality of children. The key theoretical aspect of these models is that the shadow price of children with respect to their number (i.e., the cost of an additional child, holding their quality constant) is greater the higher the quality is. Vice versa, the shadow price of children with respect to their quality (i.e., the cost of an increase in quality

holding quantity unchanged) is greater the number of children is. An important consequence is that if the "price" of a child increases (for example, due to a raise in the cost opportunity of time for the mother or due to greater investments in education) then the effect will be a substitution of quantity with quality. These theoretical considerations solicit to put great attention on education when we analyse the fertility-poverty relationship. Education, as said in the previous two sections, is a determinant of both phenomena. Moreover, different levels of parents' education imply a different consideration of children.

It is important to mention that the traditional micro-economic framework considers children as an essential part of the household's work force as they generate income, as well as providing insurance against old age. This is especially true for male children. In rural underdeveloped regions of the world, which rely largely on a low level of farming technology and where households have no or little access to state benefits, this argument makes a great deal of sense (Admassie, 2002). In this setting households will have a high demand for children. The down side is that a large number of children participating in household production hamper investment in human capital (Moav, 2005). There are of course important supply side considerations to be made in this regard: rural areas in developing countries have poor access to both educational infrastructure and contraceptives, both limiting the extent couples are able to make choices about fertility outcomes (Easterlin and Crimmins, 1985).

As households attain higher levels of income and wealth, they also have fewer children, either due to a quantity-quality trade-off as suggested by Becker and Lewis (1973) or due to an increase in the opportunity cost of women earning a higher income as suggested by Willis (1973). Expansion of female education, which reduces women's willingness to give up work for childbearing, is possibly the most important driver behind increased opportunity cost and fertility decline. Consequently, fertility reduction is often seen as a direct result of increased empowerment of women through education. Educational infrastructure and educational policies are clearly important as higher compulsory childhood schooling will delay the onset of a young adult's working life, thereby reducing

child labour (Livi-Bacci 2000; Kabeer 2001). Lack of education opportunities for women reinforces social norms of women's role and position in society.

In many traditional societies, men's status depends very much on their ability to foster a large family and household heads are often considered more successful if they have many children. Such perceptions are likely to be stronger in rural areas, where, households always show a stronger gender bias in favour of boys when deciding to send kids to school. The consequence is that women's roles tend to be limited to childrearing and other household chores. With economic progress and urbanisation, however, women gain in empowerment through higher education and independence (Drovandi and Salvini, 2004). Social norms become weaker, and traditional demographic patterns fade, which is reflected by the demographic transition. Moreover, economic progress reduces labour intensive technologies, and thereby reduces the demand for child labour.

As noted by McNicoll (1997) the interpretation of the link between poverty and fertility cannot neglect the institutional settings. Households' fertility behaviour adjusts to changes in perceived and actual cost and benefits of children. Economic forces, social organizations and cultural patterns strongly influence prices that determine costs and benefits of children. Factors like the educational system and infrastructures, health facilities, family planning policies and centres, culture, religion, social norms are all crucially important for both fertility and poverty and for the relationship between them.

Existing research on the relationship between poverty and fertility in LDC are mainly based on cross-sectional data. For a review of this literature we can refer to Schoumaker and Tabutin (1999). The results vary considerably: some studies find a negative relationship between poverty and fertility; in others the relationship seems to be very weak; in the majority of cases the relationship is found to be positive. These mixed results are explained in consideration of the different level of country development and demographic transition.

Within the poorest countries for example, the relationship between poverty and fertility is often negative. Fertility appears higher among "wealthier" households, which is a result of low reproduction capability and general higher rates of infertility among the poor (Lipton 1998; Livi-Bacci and di Santis 1998).

In some cases, such as rural areas of India and Cameroon where fertility rates are very high, the relationship takes the inverse "J shape", implying that both low and high-income households have lower rates of fertility, whereas medium level income households have higher fertility (Schoumaker and Tabutin, 1999). It is argued that very low income households tend to be landless farmers, hence less reliant on children as cheap labour, whereas those with the highest income has lower fertility due to higher investment in child quality. The middle income families are landholding farms which depend on cheap labour, and therefore have a higher demand for child quantity, which explains the apparent inverse J-shape.

The most common relationship between poverty and fertility in contemporary less developed countries is however positive. For instance countries with low fertility levels during the eighties and the nineties (TFR less than 3.5 – including Vietnam, Costa Rica, urban Paraguay, and urban South Africa) and with high fertility levels (TFR above 4.5, e.g. Guatemala, Cameroon, Bolivia, Calcutta in India, Belize), as well as medium level fertility (TFR between 3.5 and 4.5, e.g. Mexico, rural India, rural South Africa, Brazil, El Salvador, Ecuador, Paraguay), all show a positive relationship.

All of the studies referred to above are based on cross-sectional data, and as far as we are aware none have looked at the relationship in a dynamic perspective. However, with the emergence of longitudinal data, research on poverty dynamics for developing countries is now rising, though emphasis on fertility is still limited (Aassve et al, 2006b).

The project "Poverty dynamics and fertility in developing countries", involving the participation of ISER, the Vienna Institute of Demography (VID) and the Department of Statistics of the University of Florence, was developed with the purpose of compensate the gap existing in the demographic and development economics applied literature. The study of the Vietnamese case belongs to the project's goals.

The longitudinal dimension of the data available is crucially important to be allowed to draw robust causal inference about the effect of interest. In fact, only longitudinal data allow to keep into account the dynamic nature of fertility

and poverty processes. As we better explain in chapter 2, using data on two time points we are allowed to properly implement a pre-post treatment study which is vital for our study of causal inference.

1.5 The Vietnam Living Standard Measurement Survey and the Vietnamese context

As formerly said, the data we use in our applications come from the Vietnamese Living Standard Measurement Surveys (VLSMS; see for details GSO, 1994 and 2000). The first VLSMS was conducted in 1992-93 by the State Planning Committee of Vietnam (now called Ministry of Planning and Investment) along with the General Statistical Office (GSO). The second VLSMS was conducted by the GSO in 1997-98. Both VLSMS surveys were funded by UNDP and Swedish International Development Authority. The survey was part of the Living Standards Measurement Study (LSMS) household surveys conducted in a number of developing countries with technical assistance from the World Bank.

The second VLSMS was designed to provide an up-to-date source of data on households to be used in policy design, monitoring of living standards and evaluation of policies and programs. The timing of the second VLSMS approximately five years after the first allows analysis of medium term trends in living standards as a large part of the questionnaire is the same in both surveys.

Likewise all LSMS, the Vietnamese surveys include rich information on variables that are important determinants for the household's standard of living and fertility behaviour. For example, it collects data on education, employment, fertility and marital histories, together with detailed information on household income and consumption expenditure. According to Falaris (2003), the overall quality of the panel is impressive with a very low attrition rate.

A very interesting feature of the VLSMS is that it also provides detailed community information from a separate community questionnaire. Community level information is available for rural areas only and includes 120 communities, with information on markets, roads, electricity and other important

infrastructures and main economic activities. The communities in Vietnam range in size from 8,000 inhabitants to 30,000 and represent a key geographical dimension for economic, fertility and social behaviours in general.

The target sample size selected for the 1997-98 VLSMS was 6000 households. The majority of the sample was comprised of the households interviewed in 1992-93 with the first VLSS survey (4800 households). Households are defined as people living and eating meals together in the same dwelling. In most cases there is only one household per dwelling as people who live together usually eat together.

The sample in 1992-93 was a self-weighted sample drawn from all areas of Vietnam. The overall sampling frame was stratified into two groups urban and rural, with sampling carried out separately in each group (strata). According to the 1989 census, about 20% of Vietnamese households lived in urban areas so the sample stratification ensured that 20% of selected households also came from urban areas. The selection of communes was done to ensure that they were spread out evenly among all provinces in Vietnam.

The sample was drawn in three stages with communes (in rural areas) and small towns (in urban areas) chosen as the primary sampling unit as that was the lowest administrative unit for which the GSO had estimates of population in 1992. A total of 120 communes and 30 towns were out of the 10,000 in all of Vietnam with probability of selection proportional to their population size. As some communes are quite large in size, logistically it would have been difficult to interview 32 households selected randomly within each commune/ward. Instead, population figures for each village (in rural areas) or block (in urban areas) were compiled from the selected communes to select two villages/blocks randomly with probability proportional to their population size. Finally, the third stage involved listing all households within each selected village/block and selecting 20 households (16 for the sample and 4 extras if it became necessary to replace a selected household).

For the sample to reach 6000 households in 1998, an additional 1200 households were required. This was done by selecting households from the total

sample of the 1995 Multi-Purpose Household Survey of the GSO. Obviously we used only the panel households in the analysis.

The introduction of the VLSMS has sparked several poverty studies (examples include Haughton *et al*, 2001; Glewwe *et al*, 2002; White and Masset, 2002 and 2003; Justino and Litchfield, 2004). These studies suggest that female headed households, lack of education, rural households (or living in the Northern Uplands), households dependent upon agriculture, are associated with higher poverty. They also suggest that children, despite declining fertility rates, remain an important driver behind poverty. This is confirmed by Pudney and Aassve (2007ab) who shows that childbearing is strongly associated with lower living standards also during the 1990s, a period where poverty reduction was strong.

The process of poverty reduction in Vietnam started during 1980s. At the beginning of the 1980s, Vietnam was one of the worlds' poorest countries. Since then the country embarked on a remarkable recovery, a fact that is reflected by strong economic growth (Glewwe et al., 2002). The country also experienced a dramatic improvement in several indicators of social and economic wellbeing. For example, school enrolment rates increased during the period both for boys and girls. In particular, upper secondary enrolment rates increased from 6 to 27 percent for girls, and from 8 percent to 30 percent for boys (World Bank, 2000). Access to public health centres, clean water and other infrastructure have all increased, as well as the ownership of important consumer durables. Overall these improvements have had a positive effect on households' own assessment of their living conditions. As The World Bank Vietnam Development Report states: "[…] Household reports a greater sense of control over their livelihoods, reduced stress, fewer domestic and community disputes […]".

Much of this improvement has been attributed to the "Doi Moi" policy (translated in English as "renovation"). This was initiated in the late 1980s and roughly coincided with the collapse of the Soviet Union, on which Vietnam had been heavily dependent. The Doi Moi had many similarities with the reforms taking place in China a decade earlier.

The main elements of the Doi Moi were to replace collective farms by allocating land to individual households; new legalisation encouraging private

economic activity; removal of price controls; and legalisation and encouragement of Foreign Development Investment (FDI). During the nineties, immediately following the Doi Moi, Vietnam experienced a positive macroeconomic trend: strong inflation reduction, stabilization of exchange rate, sustained economic growth. The average annual GDP growth was at a staggering 7 percent. In the period covered by the Vietnam LSMS panel (i.e. from 1993 to 1998), the growth rate was even higher at 8.9 percent. This was followed by significant changes in the labour market; during the 1990s the employment grew by 2.5%. Output was increased through improved productivity and prices rose as a result of expansion in export of rice. By mid 1990s Vietnam passed to be a net importer to be one of world's largest exporters of rice on the international markets. The increase in agriculture diversification was another remarkable factor of the economic change.

Given such a strong economic performance, it is not unexpected that the overall poverty rate fell. The official poverty rate, which is derived from the per capita household consumption expenditure, declined from 58% in 1993 to 37% in 1998. Though the exact number is contested, as this depends on how poverty is measured through the equivalence scale, (Justino and Litchfield, 2004; White and Masset, 2003; World Bank, 2000), there is little doubt that poverty did indeed decline during this period.

However, economic growth by itself is not a sufficient condition for poverty reduction (Huong *et al*, 2003; Ghura *et al*, 2002; Glewwe *et al*, 2002; Bruno *et al*, 1999) and the way in which individuals may gain from the growth depends on their individual skills, education and health, ethnicity, their religion, geographical location, and type of employment and occupations. Whereas the economic boom in Vietnam affected all geographical, ethnic, and socio-economic groups, it did so in very different ways, and the poverty reduction was certainly not uniform across the population (Justino and Litchfield, 2004; Balisacan *et al*, 2003; Glewwe *et al*, 2002). In particular, it is noted that inequality increased during the nineties (Haughton et al, 2001), a fact that is robust to how inequality is measured. Gains from economic growth was stronger

in urban areas, for South East and Red River Delta[2], for Kinh[3] which is the main ethnic group in Vietnam, for households headed by a white collar worker and for those with higher education. However, the data also shows much stronger heterogeneity in poverty reduction in rural areas. There is in other words a significant degree of clustering across rural areas.

As a result we focus our analysis on the rural areas of Vietnam. Focusing on rural household has also other motivations: only the rural sample of the VLSMS contains some interesting community level information. Finally, our focus on rural households is further justified by the fact that the majority of the Vietnamese population lives in rural areas, the poorest part of the country, dominated by agriculture.

From a demographic point of view, an important aspect to bear in mind analysing Vietnam situation is that this country has experienced a tremendous decline in fertility over the past three decades, and at present one can safely claim that the country has completed the fertility transition.

The figures speak for themselves: in 1980 Total fertility Rate (TFR) was 5.0, in 2003 it was 1.9. Naturally, fertility levels in rural areas remain higher than in urban areas, but with a rural population of 80 percent, the overall TFR reflects in any case a substantial decline in fertility. Vietnam's TFR is now one of the lowest in the developing world, higher only than Thailand and China (Haughton et al, 2001).

Duy et al (2001), argue that the drop in fertility is due in about equal measure to later and fewer marriages, and to an increase in contraceptive use. The proportion of married women who say they are using modern contraceptive methods, particularly IUDs, is very high, having risen from 43.9% in 1993 to 55.1% by 1998. Contraceptive use rates also vary less across regions than they did in 1993; the Mekong Delta in particular has largely closed the contraceptive

[2] The Red River Delta and the Mekong River Delta were the regions that benefited more from rice market liberalisation (Justino and Litchfield, 2004).
[3] In Vietnam there is a large population of ethnic minorities that tend to be significantly poorer than Kinh majority. An analysis of the sources of the ethnic inequalities in Vietnam is found in Van de Walle and Gunewardana, 2001.

use gap with the rest of the country (Haughton et al, 2001; Anh and Thang, 2002).

However, the previous considerations do not clarify what fundamental forces are behind the drop in fertility. Bearing in mind the theoretical economic considerations discussed in section 1.4, we can argue that Vietnamese households are moving from a desire for a large quantity of children to a preference for quality, but this begs the question of why such a shift is underway. Possibly the mixture of rising and high educational costs along with reduced labour contributions from children (who are more likely to be at school) and changed expectations about how to finance old age, may be combining to make having children less attractive (Haughton et al, 2001). Increasing urbanization, high and rising levels of maternal education, and a vigorous family planning program also play a part.

1.6 Why adopting a causal and a multilevel perspective in studying the relationship between fertility and poverty?

In the previous sections we outlined the most important theoretical considerations, to be kept into account in the analysis of the fertility-poverty relationship. In this section we want to highlight two key aspects for the development of the following chapters. In particular, we want to focus the attention on two approaches we use in the thesis: causal and multilevel.

We are interested in causality when we want to take a decision. Policy-makers are naturally interested in causality. Good public policy decisions require reliable information about the causal relationships among variables. Policy-makers must understand the way the world works and the likely effects of manipulating the variables that are under their control. If, for example, having more children causes poverty, policy makers could plan, adequately, some actions to impact on fertility, directly or indirectly. Alternatively, if the only policy goal is to contrast poverty conditions without any wish to determine

fertility behaviours, it could be simply decided to compensate the higher costs supported by households with a lot of children through state benefits.

A complication in the study of the causal relationship between fertility and poverty is the reverse causality problem: theoretically, not only fertility impacts on poverty but also the contrary could be true. However, previous researchers have found little or no effect of fertility on poverty (e.g., Aassve et al, 2006ab). Also in our application we found some evidence against a causal feedback from fertility to poverty. In the thesis, we hence concentrate on the impact of fertility on poverty. Moreover, we think that also theoretically fertility *per se* cannot *impact* directly on poverty. In the next chapter we discuss the approach to causal inference we adopt and some methods for the estimation of causal effects in observational studies.

As the multilevel perspective is concerned, this is motivated by the consideration that the place where a household reside is important, as we already argued, both for poverty and fertility. In particular, households can be considered as clustered in communities. This implies a two-level data structure, with households at the first level and communities at the second.

Keeping explicitly into account this multilevel dimension in the study of the causal effect of fertility on poverty is important, both for statistical reasons and for substantive research questions. The fact that community characteristics (infrastructure, remoteness, culture and so on) influence both phenomena requires to control also for them in the statistical analyses. Otherwise, we risk to find spurious effects. Moreover, the multilevel dimension implies specific challenges for causal inference that we discuss extensively in chapter 3.

Apart from the statistical motivations, the multilevel structure of the data brings some interesting research question. For example, it is of interest to understand if the effect of fertility on poverty changes by community.

Community characteristics can determine a considerable heterogeneity in the effect of fertility on poverty for several reasons. For example, the presence of some specific facilities in the community may help women (and families) to rise up children. Examples could be health facility centres which could offer sanitary assistance (eventually free or partly free). Also the quality of facilities is

important and it varies considerably in Vietnam by province, district and even commune (Evans et al, 2007).

In Vietnamese rural communities, as in all rural areas in LDC, also unofficial forms of "assistance" can be very important. As noted by Justino (2005), "the most common forms of social security in Vietnam, as in most developing countries, are informal and delivered through family and community social networks". This informal social security system includes informal work exchange, food assistance among neighbours, loans made available by family and moneylenders.

The general economic community environment is also important. For example, if a community shows a high degree of economic development it is more likely that women loosing work can easily find another work after the first years of life of children. Finally, community institutional setting and social norms are crucial. It could be the case that the social system and the family taxation legislation are defined at national level but at the local level some forms of autonomy is leaved. For example, communities could have the power to assign some benefits to families with more children or to the new born children. This seems not to be the case in Vietnam. In fact, income taxation play a very small role in Vietnamese fiscal policy. However, some considerations are important. The Vietnam taxation system ensures a certain degree of autonomy to provinces. But we were not able to know on what this autonomy precisely consists and then we are not able to know if it can be an effect on the impact of childbearing events. Apart from this, however, at local level a series of charges for local environmental policies, school enrolment and other services exist. The amount of these charges varies considerably by community. In addition to formal charge, there are also informal charges and gifts due to local public servants for any services (Evans et al, 2007).

Finally we note that Vietnam social security system envisages a maternity benefit (Evans et al, 2007; Justino, 2005; S.S.A, 2006). As noted by Justino (2005) in areas such rural Vietnamese communities it is much likely that the concrete intake of these kind of provisions heavily depends on the context, and specifically on the competences, skills or training of the municipality employees,

isolation of the community, global education (in more educated communities is more likely to know about benefits and how to get them). Administrative inefficiency and low literacy levels can prevent people from claiming the benefits to which they are entitled. These aspects are discussed, from a methodological point, of view in chapter 3, while in chapter 6 their importance is assessed in the context of our application.

1.7 Concluding remarks

This chapter, basically, serves as a base for the following discussion. We have introduced the concept and measure of poverty we will use in our analyses and discussed the main determinants of poverty and fertility. Understanding the common determinants of the two phenomena is crucial in order to address the key question of the thesis: is there a causal effect of fertility on poverty?

The determinants of poverty and fertility can be distinguished in three groups: individual, household and aggregate. Individual and household determinants include demographic, economic and social characteristics. Aggregate determinants refer to characteristics of the place where people reside: infrastructures, social norms, institutions, and so on.

As emerging from theoretical and empirical researches, education plays undoubtedly, a key role for fertility behavior, as well as for living standards conditions. Expansion of female education, which reduces women's willingness to give up work for childbearing, is possibly the most important driver behind increased opportunity cost and fertility decline. Consequently, fertility reduction is often seen as a direct result of increased empowerment of women through education.

Analyzing the literature about the relationship between fertility and poverty, we noted that few works use longitudinal data and a causal perspective. We have explained that adopting a causal point of view is vital to give valuable policy advices. The availability of longitudinal data is a valuable aspect to give robustness to the estimation of causal effects.

The other fundamental approach we take in the thesis is the multilevel one. This is justified by the consideration that both fertility and poverty are heavily influenced by the characteristics of the community where households live. In this perspective, recognizing this two-level structure (household within communities) is important for statistical and substantive issues, as we more meticulously explain in chapter 3.

Chapter 2
Causal inference in observational studies under the potential outcomes framework

Introduction

In this chapter we present the framework in which our causal analysis is developed and compare several approaches to handle the issues of selection bias and endogeneity. Whereas we develop our reasoning in the light of the specific context of our application, we present here a more general framework for causal inference.

We start the chapter (section 2.1) by discussing the potential outcomes framework that we adopt in this work. We introduce key concepts, such as assignment mechanism, randomised experiment and observational study. We then discuss the estimation of causal effects in randomised (section 2.2) and observational studies (sections 2.3 and 2.4). Since our application concern an observational study, we focus on this kind of setting. In section 2.3 we present methods for estimating causal effects in those situations where the unknown assignment mechanism is assumed to be regular. In this setting we contrast regression and propensity score matching methods, with the goal to highlight differences and similarities among these methods. Section 2.4 analyses studies of causal inference in which the assignment to treatment is assumed to be regulated by a latent regular mechanism. In this context, we explore the use of instrumental variables for the identification of causal effects. We distinguish randomised and conditionally randomised instruments, where the difference lay on the fact that

the latter can be assumed to be randomised only conditional to a set of observed covariates.

2.1 The potential outcomes framework

In many field of social sciences there is a growing interest in the methods that can be used to evaluate the effects of social programs and public policies. While academic researchers are increasingly focused on assessing the strengths and weaknesses of the evaluation methodology itself, policy makers are turning to the results to provide the foundations for evidence-based policy.

Causal inference is a part of the statistical research which has received increasing interest in recent decades. In this work we adopt the counterfactual or potential outcomes framework to causal inference, which was pioneered by Neyman (1923) and Fisher (1925) and extended by Rubin (1974, 1978) to observational studies. Recently the approach has been adopted by many in both statistics and econometrics (e.g. Rosenmbaum and Rubin, 1983; Heckman, 1992 and 1997; Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Heckman et al, 1997). This literature formalise notions of cause and effect and is based on the counterfactual idea. Counterfactual refers to what would have happened if, contrary to fact, the exposure had been something other than what it actually was (Greenland and Brumback, 2002).

Two fundamental assumptions of the standard potential outcomes framework are that:

a) each unit in the population of interest could have received any one of the treatments,
b) for each unit *i* and treatment *d*, at the time of treatment assignment, the outcome that individual *i* would have if the unit gets treatment level *d* exists, even if the individual does not in fact get *d*; this value is called the *potential outcome* of the unit under treatment *d*.

Let us suppose we have a population of individual units under study (in our case households) indexed by $i = 1, 2, \ldots, N$, an indicator for a binary treatment[1], D, which assumes the value 1 for treated units and 0 for untreated, or controls, and an outcome variable, which we indicate by Y. Each unit, i, has two potential outcomes depending on the assignment to the treatment levels: Y_{i1} if D_i=1 and Y_{i0} if D_i=0. Potential outcomes for unit i and treatment d can be written as Y_{id}, with $d \in \{0,1\}$. The fact that this variable is labelled only by i and d corresponds to the "no interference among units" assumption of Cox (1958), which Rubin (1980) refers to as the Stable Unit Treatment Value Assumption (SUTVA).

SUTVA consists of two components. The first states that the potential outcomes for any unit do not vary with the treatments assigned to any other units. In our application it means that having a child has an effect on household consumption, independently of fertility behaviours of the other households. The second component requires that there are no versions of the treatment (and controls) (i.e. treatment characteristics are the same for each treated (control) units). In our application this implies that the characteristics of the new child (sex, weight, etc) are not relevant. Without this assumptions we would no longer have only two potential outcomes for unit i.

As with many of the other assumptions to be discussed, it is important to note that SUTVA is not directly informed by the data. In other words, it is an untestable assumption that stems from the scientist's assessment or knowledge. SUTVA has been criticized in many situations, an example being the evaluation of the effect of a vaccination campaign on a contagious disease (Halloran and Struchiner, 1995). The authors are worried about the fact that interference among units exists. In the context of this application, in fact, it is plausible that the effect of vaccinating a unit changes according to the number of the other units being vaccinated. Another context in which the assumption is criticized is when evaluation is undertaken within multilevel structures (Subramanian, 2001), an example being educational research, where students are clustered in schools.

[1] In this work we use binary treatments. Recently, the literature has extended the potential outcome framework to cases of multi-valued (Imbens, 2000; Lechenr, 2001), as well as continuous treatments (Imai and van Dyk, 2003; Imbens and Hirano, 2004).

Whereas several approaches have been developed to allow for violation of the SUTVA, we maintain this assumption in the following discussion. We discuss possible sources of violation, and resulting methods to deal with this problem in the light of our application in chapter 3.

Under the SUTVA, the N x 2 matrix of all potential responses defines the true population: $\boldsymbol{Y} = (\boldsymbol{Y_0}, \boldsymbol{Y_1})$, where $\boldsymbol{Y_d}$ is the N x 1 vector of responses if all the units in the population were to receive treatment d. Following Rubin (1978) the true causal effect of the treatment for a given unit i, is defined as the comparison between the two potential outcomes Y_{i1} and Y_{i0}, which can be constructed, for example, as a ratio or a difference. If we consider simply a difference, the true casual effect can be written, for the unit i as:

$$\Delta_i = Y_{i1} - Y_{i0}. \tag{2.1}$$

It is obvious that the two potential outcomes in (2.1) are not observable for the same unit at the same time. Holland (1986) refers to this aspect as the "fundamental problem of causal inference". For each unit we can observe only one of the two potential outcomes, according to the treatment the unit actually received, the other being missing. In this sense, causal inference can be considered as a missing data problem.

The following relationship makes clear that the observed outcome for unit i, which we indicate with Y_i^{obs} depends on the treatment indicator

$$Y_i^{obs} \equiv Y(D_i) = D_i * Y_{i1} + (1-D_i) * Y_{i0} = Y_{i0} + D_i * (Y_{i1}-Y_{i0}). \tag{2.2}$$

The last equality in (2.2) states that the treatment "adds" a quantity (Y_{i1}-Y_{i0}) to the outcome with respect to the case of no treatment. We indicate with $\boldsymbol{Y^{obs}}$ the N x 1 vector of outcomes observed on the N units under study.

In spite of the apparent non resolvability of the "fundamental problem", several approaches are developed to overcome it. One option is to observe the same physical unit at different points in time. Under certain strong assumptions,

the individual specific causal effect (that is, the causal effect that refers exclusively to that unit) will be identified. However, one would need to assume *temporal stability,* in the sense that the effect of the treatment does not depend on when it is applied, as well as *causal transience*, which imply that the exposure to control (treatment) at time *t* does not affect the result of the exposure to treatment (control) at succeeding times (see Holland, 1986).

However, in most cases we are not interested to the estimation of a causal effect on a single unit, but on the entire population, or on a relevant sub-group of it. Hence, the statistical solution to the fundamental causality problem consists to substitute the estimate of individual causal effects by an estimate of an average causal effect. Therefore, in order to make causal inference possible, a key requirement is that of *replication* (see Stuart, 2007).

Replication means, in this context, that there must be multiple units for which we can observe each one of the potential outcomes. If only one unit was assigned either to treatment or control, we would have no sufficient information. However, with some units receiving the treatment and others the control, we can use the treated units to learn about the potential outcomes under treatment (that is, for control units we can "impute" a potential outcome under treatment using information on observed outcomes for treated), and the control units to learn about the potential outcomes under control. In this regard, we can argue that potential outcomes models have the main virtue, with respect to other approaches, of make explicit the need of what Maldonado and Greenland (2002) call a *substitution step*, which refers to the need of estimate a counterfactual outcome: for treated units we need to estimate what would have been their outcome if they were exposed to control, and *vice versa* for controls. The problem attributed to modelling such unobserved quantities, which in essence implies the need for an un-testable assumption, is part of the intrinsic problem of causal inference. Whereas this issue comes up-front in this framework, it is often obscured in alternative approaches (Greenland and Brumback, 2002).

An important insight from the potential outcomes approach is that the presence of multiple units does not solve the problem of causal inference. Rather, multiple units guarantee that associational parameters can be calculated,

but for the identification of causal effects we always need to impose some assumptions, whose plausibility depends on the specific application under study.

The potential outcomes framework is not the only approach to causal inference that has been developed in the literature. Other important approaches are represented by Structural Equation Models (SEM) and Direct Acyclic Graphs (DAG).

Structural Equation Models (SEM) have a long history, dating back to path analysis developed by geneticists (Wright, 1928)[2]. Structural equations methods are predominant in economics for modelling, identifying, and estimating causal effects of interest. The early work of the Cowles Commission studied identification and estimation of causal effects (e.g. Haavelmo, 1943; Simon, 1954). This literature also introduced notions of "endogeneity" and "exogeneity." Reiersøl (1945) formalized the method of "instrumental variables" (IV), originally introduced by Philip Wright (1928), within the structural equations framework. Ever since, this method has played a central role in handling issues of endogeneity (e.g. Goldeberger, 1972; Heckman, 1997; Angrist and Krueger, 2001). Structural equation models, basically, rely on the specification of systems of equations with parameters and variables that attempt to capture behavioural relationships and specify the causal links between variables. Inference in SEM often exploits the presence of some instrumental variables, which are variables explicitly excluded from some equations and included in others. As we will see in section 2.4, Angrist et al (1996) provide a link between the potential outcomes and the SEM approach.

Another line of research has emerged in the machine learning literature in the work of Pearl (1995, 2000), Spirtes, Glymour, and Scheines (1993), and Dawid (2002) among others. In particular, Pearl (1995) introduced two methods related to the labor economics and treatment effect literatures, the "back door" and "front door" methods. A distinctive feature of this literature is the use of

[2] A curiosity is reported by Stock and Trebbi (2003) about the authorship of the solution of the identification problem in a system of simultaneous equations. The earliest know work on this issue is included in the appendix of the book written by Philip G. Wright, *The tariff on animal and vegetable oils*. Because this appendix differ so from the rest of the book, its authorship has been in doubt. There is, in fact, another plausible author: Philip G. Wright's eldest son, Sewall. Stock and Trebbi implemented a stylometric analysis, which evidence favours Philip G. Wright.

directed acyclic graphs (DAGs) to represent causal relations and of graphical criteria to determine if particular causal effects are identifiable, with less attention to the estimation of these causal effects. Recently, White and Chalak (2006) propose the "settable system" framework as a means to unify all the three cited approaches.

In the sequel we will see how by using the potential outcomes approach we are able to define several causal parameters of potential interest and, at the same time, make clear the assumptions needed to estimate them. We do so by demonstrating different methods. Our aim is to present the different methods in a comparative way and make explicit the differences between them with respect to the underlying assumptions and data requirement. We present these methods in sections 2.3 and 2.4. In the next section we analyse some fundamental aspects for causal inference, such as the role of covariates, which are fundamental for the following discussion.

2.1.1 The role of covariates

In most empirical studies concerned with causal inference, researchers have information about other variables than just the treatment indicator, D, and the outcome, Y. These are normally termed *attributes* or *covariates* and represent characteristics that the unit possesses before exposure to the treatment. For this reason they are called also *background* or *pre-treatment* characteristics. Supposing that we have M covariates, these can be collected into a N x M matrix that we indicate, from now on, with $\boldsymbol{X}$.

The role of these variables in causal inference is threefold. As covariates do in many statistical applications, these attributes serve to make analysis more precise by controlling for parts of the variation in the outcome. Second, for more substantive reasons, the researcher may be interested in causal effects for subpopulations defined by values of these variables. This is the case when researchers suspect that some heterogeneity is present in the treatment effect, and it is of substantial interest to explore which characteristics drives it. However, the most important role of covariates in causal inference concerns their effect on

the assignment to the treatment. Often, assumptions about the assignment mechanism are more plausible if made within homogeneous subpopulations than in the overall population. In other words, some assumptions are valid only conditioning on these covariates. It is critically important that these variables are not affected by the treatment. Rosenbaum (1984) examines the consequences of including potentially predetermined variables in the estimation and concludes that such adjustment results in unbiased estimates only if the variables are not affected by the treatment. Often, the covariates take their values prior to the unit being exposed to the treatment and, in this case, it is natural to think that they are not influenced by the treatment exposure. However, this is not sufficient for the conditions they need to satisfy. In fact, it could be the case that units have some expectation about the future values of the outcome and it is possible that the decision to take the treatment can affect these covariates through some "anticipation effect".

As we will discuss in the section 2.3 and in chapter 5, in observational studies, as we will see, it is of crucial importance, in order to draw correct causal inference, to have a sufficient rich set of observed covariates, since it make more plausible the assumptions on which researcher relies. We discuss these aspects in more details in the following.

2.1.2 Causal parameters of interest

Under the potential outcomes framework we can define several causal parameters, but the ones receiving most attention in the literature are the Average Treatment Effect (ATE), the Average Treatment Effect on the Treated (ATT), and the Average Treatment Effect on the Untreated (ATU). They are defined as follows:

$$\text{ATE} = E(Y_{i1}-Y_{i0}), \tag{2.3}$$

$$\text{ATT} = E(Y_{i1}-Y_{i0}|\, D_i=1), \tag{2.4}$$

$$\text{ATU} = E(Y_{i1}-Y_{i0}|\, D_i=0). \tag{2.5}$$

The ATE is the expected effect of the treatment on a randomly drawn unit from the population. The relevance of this parameter for policy analysis is often questioned, since it averages across the entire population and, hence, includes units who would be never eligible to the treatment (Heckman, 1997). For example, if a program is specifically targeted at low income individuals, there is little interest in the effect of such a program for someone being extremely well off. Heckman et al (1997) argue that the subpopulation of treated units is often of more interest than the overall population in the context of narrowly targeted programs. Therefore, the most prominent evaluation parameter is the average treatment effect on the treated (ATT), which focuses explicitly on the effects on those for whom the program is actually intended. In particular, the ATT gives the expected effect of the treatment on a randomly drawn unit from the population of treated. It is therefore more interesting for policy makers. The ATU, on the contrary, is the effect on the subpopulation of controls and is not frequently used in the evaluation literature.

It is simple to note that the ATE could be written as a weighted average of ATT and ATU:

$$\text{ATE} = \text{ATT} * P(D=1) + \text{ATU} * P(D=0), \tag{2.6}$$

where $P(D=1)$ and $P(D=0)$ are, respectively, the proportion of treated and controls units in the population. Therefore, the three parameters coincide when the average effect of the treatment is equal in the treated and controls population. It is straightforward to verify that this happens when the difference Y_1-Y_0 is independent of D or, less strongly, if it is mean independent of D. In particular, this condition holds if both potential outcomes Y_1 and Y_0 are independent, or simply mean independent, of D. In fact independence implies that:

$$E(Y_1|D=1) = E(Y_1|D=0) = E(Y_1) \tag{2.7}$$

and

$$E(Y_0|D=1) = E(Y_0|D=0) = E(Y_0). \tag{2.8}$$

From (2.7) and (2.8) follows that:

$$\text{ATT} \equiv E(Y_1\text{-}Y_0|D\text{=}1) = \text{E}(Y_1\text{-}Y_0|D\text{=}0) \equiv \text{ATU} = E(Y_1\text{-}Y_0) \equiv \text{ATE}. \qquad (2.9)$$

As we already mentioned, in many applications we would be interested to estimate the effect of the treatment on a sub-population defined on the basis of the values of one or more covariates. Therefore, it is of interest to consider also the following conditional versions of parameters (2.3)-(2.5):

$$\text{ATE}(x) = E(Y_{i1}\text{-}Y_{i0}|\ X\text{=}x), \qquad (2.10)$$

$$\text{ATT}(x) = E(Y_{i1}\text{-}Y_{i0}|\ D_i\text{=}1, X\text{=}x), \qquad (2.\ 11)$$

$$\text{ATU}(x) = E(Y_{i1}\text{-}Y_{i0}|\ D_i\text{=}0, X\text{=}x). \qquad (2.\ 12)$$

Likewise we shown in (2.6), also ATE(x) can be thought as a weighted average of ATT(x) and ATU(x):

$$\text{ATE}(x) = \text{ATT}(x) * P(D\text{=}1|\ X\text{=}x) + \text{ATU}(x) * P(D\text{=}0|\ X\text{=}x), \qquad (2.13)$$

where $P(D\text{=}1|\ X\text{=}x)$ and $P(D\text{=}1|\ X\text{=}x)$ are, respectively, the proportion of treated and controls units in the population with $X\text{=}x$. Similarly to the previous discussion, we can note that if Y_1 and Y_0 are independent of D conditional on X (or simply mean conditional independent) then ATT(x) = ATU (x) = ATE(x) for each x in the support of the variable(s) X. In fact conditional independence implies:

$$E(Y_1|D\text{=}1, X\text{=}x) = E(Y_1|D\text{=}0, X\text{=}x) = E(Y_1|\ X\text{=}x) \qquad (2.14)$$

and

$$E(Y_0|D\text{=}1, X\text{=}x) = E(Y_0|D\text{=}0, X\text{=}x) = E(Y_0|\ X\text{=}x). \qquad (2.15)$$

From (2.14) and (2.15) follows that:

$$\begin{aligned} \mathrm{ATT}(x) &\equiv E(Y_1 - Y_0 | D=1, X=x) = \mathrm{E}(Y_1 - Y_0 | D=0, X=x) \\ &\equiv \mathrm{ATU}(\mathrm{x}) = E(Y_1 - Y_0 | X=x) \equiv \mathrm{ATE}(x). \end{aligned} \tag{2.16}$$

The marginal average effects can be obtained from the respective conditional versions by averaging with respect to the distribution of X on the appropriate population. For example, the ATE can be obtained from the ATE(x) by averaging over the entire population:

$$\mathrm{ATE} = E_X[\ \mathrm{ATE}(x)\ ] = \int \mathrm{ATE}(x)\ f_X\ dx\,, \tag{2.17}$$

where f_X is the density function of X in the w hole population. In a similar way, the ATT and ATU are obtained as follows :

$$\mathrm{ATT} = E_{X|D=1}[\ \mathrm{ATT}(x)\ ] = \int \mathrm{ATT}(x)\ f_{X|D=1}\ dx\,, \tag{2.18}$$

$$\mathrm{ATU} = E_{X|D=0}[\ \mathrm{ATU}(x)\ ] = \int \mathrm{ATU}(x)\ f_{X|D=0}\ dx\,, \tag{2.19}$$

where $f_{X|D=1}$ and $f_{X|D=0}$ represent, respectively, the density function of X in the treated and in the controls population. Let us consider the three important cases:

1) ATE(x)=ATT(x)=ATU(x)=r(x) for each x,
2) ATE(x)=ATT(x)=ATU(x)=r(x)=λ for each x,
3) $f_{X|D=1} = f_{X|D=0}$,

where r(x) is a real-valued function and λ is a constant.

In the first case, the three conditional parameters are equal for each value of the covariate X, but they are allowed to vary by the X-values. In this case, the marginal parameters (ATE, ATT and ATU) are, in general, different. In fact, from (2.17), (2.18) and (2.19) it is easy to see that they weight in different ways the values r(x). In the second case, instead, ATT, ATU and ATE coincide.

Finally, under the third condition the three marginal parameters will coincide even if r(x) is not a constant function (first case). From this discussion,

we learn that the first and the third conditions combined, or the second condition taken alone, are sufficient for the equality of ATE, ATT and ATU. On the other hand, in case of heterogeneous treatment effects *and* different distribution of covariates in the treated and control population, the three parameters are expected to be different[3]. We will use this discussion in the interpretation of results of our application in chapter 5.

Before discussing methods for the estimation of the parameters of interest we need to introduce a key concept in causal inference: the assignment mechanism.

2.1.3 The assignment mechanism

In the previous section, in order to define the causal parameters we only used the potential outcomes definition and SUTVA. Importantly, the causal parameters are defined independently of which potential outcomes we actually observe. It is the fact that we do not observe all potential outcomes that induces inferential problems, in which we need to rely on statistical techniques (Imbens, 2007). In this sense, we already said, that the problem of causal inference is a missing data problem. A key issue in the missing data literature is the so called missing data mechanism, which in the causal inference framework is often termed the assignment mechanism (Rubin, 1978). Its role is fundamental in the sense that we cannot draw valid causal conclusions without considering what makes some units receive a treatment, whereas others do not.

In simple terms, the assignment mechanism is defined as the mechanism that determines which units get which treatment. More formally, the mechanism is defined as a function that assigns probabilities to all possible N x 1 vectors of

[3] Obviously, in an observational study both conditions 2 and 3 are rarely respected. We note, for completeness, that by chance ATT and ATU, and hence ATE, can coincide. For example, let suppose we have only one covariate X taking three values: $X = \{1,2,3\}$. And that ATT(X=1)=300; ATT(X=2)=100; ATT(X=3)=100; ATU(X=1)=100; ATU(X=2)=200; ATU(X=3)=300. Finally, let assume that the conditional distribution of X are: $P(X=1|D=1)=10\%$; $P(X=2|D=1)=50\%$; $P(X=3|D=1)=40\%$; $P(X=1|D=0)=50\%$; $P(X=2|D=0)=40\%$; $P(X=3|D=0)=10\%$. Then, it easy to verify that ATT=ATU=ATE=160. However, this is an extreme situation and, in general, when a difference between ATE and ATT is found this is a sign for heterogeneous effects and for unbalance in the distribution of covariates in the treated and control groups.

binary assignments $\boldsymbol{D}$ given the N x 1 vectors of potential outcomes $\boldsymbol{Y_1}$ and $\boldsymbol{Y_0}$ and the N x M matrix of covariates $\boldsymbol{X}$. The notion of assignment mechanism is further formalized in the following definition:

Definition 2.1 - Assignment mechanism

Given a population of N units, the assignment mechanism is a row-exchangeable function $Pr(\boldsymbol{D; X, Y_1, Y_0})$ taking on values in $\{0, 1\}^N$ satisfying

$$\sum_{\boldsymbol{D}} \Pr(\boldsymbol{D; X, Y_1, Y_0}) = 1 , \tag{2.19}$$

for all $\boldsymbol{X, Y_1, Y_0}$.

The probability assignment to the treatment for individual units can be defined according to definition 2.2:

Definition 2.2 - Unit assignment probabilities

The unit assignment probability for unit i is

$$Pr_i(D_i=1| \mathbf{X, Y_1, Y_0}) = \sum_{\boldsymbol{D}|D_i=1} \Pr(\boldsymbol{D; X, Y_1, Y_0}) \tag{2.20}$$

We can rewrite (2.20), for convenience, distinguishing element of $\boldsymbol{X, Y_1, Y_0}$ relative to i from those concerning the other N-1 units:

$$Pr_i(D_i=1| \mathbf{X, Y_1, Y_0}) = q(X_i, Y_{i1}, Y_{i0}, \boldsymbol{X_{-i}, Y_{-i1}, Y_{-i0}}) \tag{2.21}$$

where $\boldsymbol{X_{-i}, Y_{-i1}, Y_{-i0}}$ are obtained, respectively, deleting *ith* row from $\boldsymbol{X}$ and *ith* elements from $\boldsymbol{Y_1, Y_0}$. The (2.21) it is interesting since, depending on the

mechanism operating in a specific application, the function $q(\cdot)$ will be free of dependence on some of its arguments.

When considering the assignment mechanism, there is an important distinction between randomized and non-randomized experiments (or observational studies). The key difference is that in randomized setting the researcher can control assignment to treatment and the probabilities of treatment are known. In observational studies these conditions are unlikely to hold and the researcher can only estimates probabilities of assignment to treatment on the basis of the data available. In terms of inference, the simplest possible assignment mechanism (and the one that traditionally has been viewed as the only credible base of causal inference) is randomized assignment. It is defined as follows:

Definition 2.3 - Randomized experiment

A randomized experiment is an assignment mechanism that is

1) *probabilistic*, that is unit assignment probabilities lies strictly between 0 and 1 for each unit;
2) *ignorable*, which means that it does not depend on the unobserved potential outcomes;
3) a known function of its arguments.

An important special case is the *completely randomized experiment*. This is defined so that the assignment mechanism is *locally independent*. This means that the assignment mechanism is separable in the unit assignment probabilities, at least conditional on ($\boldsymbol{D}$, $\boldsymbol{X}$). We can thus write:

$$\Pr(D|X,Y_1,Y_0) =$$

$$g(D;X)\prod_{i=1}^{N}\Pr_i(D_i=1\mid X,Y_1,Y_0)^{W_i}\left(1-\Pr_i(D_i=1\mid X,Y_1,Y_0)\right)^{1-W_i} \qquad (2.22)$$

Another condition for the assignment to be locally independent is that the individual assignment probability function $q(\cdot)$ introduced in equation (2.21) has to be free of $\boldsymbol{X}_{-i}$, $\boldsymbol{Y}_{-i1}$, $\boldsymbol{Y}_{-i0}$, i.e., $Pr_i(D_i=1|\ \boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_0}) = q(X_i, Y_{i1}, Y_{i0})$, for all i and for some function $q(\cdot)$. An important characteristic of completely randomized experiments is that it is said to be *unconfounded*, which in essence means that it is independent of the potential outcomes:

Definition 2.4 - Unconfounded assignment mechanism

An assignment mechanism is unconfounded if the assignment mechanism does not depend on the potential outcomes:

$$Pr(\boldsymbol{D};\ \boldsymbol{X},\ \boldsymbol{Y_1},\ \boldsymbol{Y_0}) = Pr(\boldsymbol{D};\ \boldsymbol{X},\ \boldsymbol{Y'_1},\ \boldsymbol{Y'_0}), \tag{2.23}$$

for all $\boldsymbol{X}, \boldsymbol{Y_1}, \boldsymbol{Y_0}, \boldsymbol{Y'_1}, \boldsymbol{Y'_0}$.

It is important to note that in (2.23) the independence holds conditional on $\boldsymbol{X}$ and could fail marginally. In contrast to randomized experiments, inference becomes considerably more complicated when assignment probabilities are unknown to the researcher. These situations are generally referred to as observational studies.

Definition 2.5 - Observational studies

An assignment mechanism corresponds to an observational study if the assignment mechanism is an unknown function of its arguments.

In observational studies we distinguish between *regular* and *irregular* assignment mechanisms.

Definition 2.6 - Regular assignment mechanism

An assignment mechanism is regular if it is

1) probabilistic,
2) ignorable,
3) locally independent.

Regular assignment mechanisms are in observational studies the equivalent of a classical randomized experiment. The only difference between the two types of assignments concerns the knowledge of the researcher about the assignment mechanism. Because ignorability and local independence imply unconfoundedness (for a formal proof see Imbens, 2002), regular assignment mechanisms, just as classical randomized experiments, are always unconfounded.

Irregular assignment mechanisms represent the most difficult situation and for these mechanisms there is no general approach. Our interest will be on those designs in which the assignment mechanism is assumed to be "*latent regular*", which means that it is regular given certain covariates which are not observed. Randomized experiments with non compliance, and by extension, instrumental variables analysis, belong to this setting.

In the following we present different approaches to the estimation of causal effects, organized on the basis of the assignment mechanism.

2.2 Estimating causal effects in randomised experiments

In the previous discussion we ignored the potential problem of non-compliance in randomized experiments. Non compliance takes place when the researcher cannot force units that are assigned to a specific level of the treatment to comply with this assignment (Imbens and Angrist, 1994). In other words, some units assigned to the treatment may choose to obtain the control and vice versa. These are defined as *non compliers*. In contrast, units that receive the treatment for which they are assigned are termed *compliers*. In this section we abstract from

this potential complication. The issue of non-compliance can be studied by using the instrumental variable setting we discuss in section 2.4.

Randomized experiments without non-compliance are the easiest case to treat from an inferential point of view. We can distinguish between two situations: assignments that are made independently of any unit characteristics and assignments in which probabilities depend on covariates. In the first case, randomization implies that the treatment indicator $\boldsymbol{D}$ is statistically independent of potential outcomes $\boldsymbol{Y_1}$ and $\boldsymbol{Y_0}$:

$$Y_{i1},\ Y_{i0} \perp D_i, \tag{2.24}$$

where $\perp$ in the notation introduced by Dawid (1979) means independence. As illustrated in section 2.1.2, an important consequence of (2.24) is that ATT, ATU and ATE coincide.

In these situations, randomization ensures that, on average, units belonging to the treated and comparison group differ only with respect to the treatment status and, therefore, if we observe a difference in the average outcome between the two groups this can be addressed to the treatment effect. Formally, assuming (2.24) and taking expectations of equation (2.2) yields

$$E(Y_i^{obs} \mid D_i = 1) = E(Y_{i1} \mid D_i = 1) = E(Y_{i1}),$$

and

$$E(Y_i^{obs} \mid D_i = 0) = E(Y_{i0} \mid D_i = 0) = E(Y_{i0}).$$

Thus,

$$\text{ATE } (= \text{ATT} = \text{ATU}) = \text{E}(Y_i^{obs} \mid D_i = 1) - \text{E}(Y_i^{obs} \mid D_i = 0). \tag{2.25}$$

The right-hand side of (2.25) is easily estimated through its sample equivalent, as the difference in the sample means of the observed outcomes, Y_i^{obs}, in the two groups:

$$\hat{ATE} = \frac{\sum_{i:D_i=1} Y_i^{obs}}{\#\{D_i = 1\}} - \frac{\sum_{i:D_i=0} Y_i^{obs}}{\#\{D_i = 0\}} \tag{2.26}$$

The (2.26), which is referred to also as the naïve estimator, gives an unbiased, consistent and asymptotically normal estimate of ATE (Wooldridge, 2002).

The independence assumption on which estimator in (2.26) relies is, obviously, very strong in observational studies. However, in order to estimate correctly ATE through (2.26) we need a less restrictive, but still strong condition. What we need is that independence holds, at least, on average:

$$E(Y_{i1}| D_i) = E(Y_{i1}) \text{ and } E(Y_{i0}| D_i) = E(Y_{i1}). \tag{2.27}$$

Moreover, if the parameter of interest is the ATT we need to impose only that

$$E(Y_{i0}| D_i) = E(Y_{i1}). \tag{2.28}$$

In fact, if we write the ATT as:

$$\text{ATT} = E(Y_{i1} - Y_{i0}| D_i = 1) = E(Y_{i1} | D_i = 1) - E(Y_{i0}| D_i = 1)$$

we note that the first term, $E(Y_{i1}| D_i = 1)$, can be easily estimated through its sample analog, $E(Y_i^{obs}|D_i = 1)$ and that, given (2.28), $E(Y_{i0}| D_i = 1) = E(Y_{i0}| D_i = 0)$ and hence the second term be estimated through $E(Y_i^{obs}|D_i = 0)$.

In the second group of randomized experiment, we have cited before, assignment probabilities depend on covariates. For increase efficiency, for example, researcher could randomize assignment to treatment within blocks defined by the values of certain observed covariates $\boldsymbol{X}$, which are known to influence the outcome. This type of randomization is called *stratified*

randomized experiment or *randomized blocks*. In this case, condition (2.24) holds within blocks defined by the covariate values. Hence, we can re-write it as:

$$Y_{i1},\ Y_{i0} \perp D_i \,|\boldsymbol{X_i}. \tag{2.29}$$

Whereas independence does not hold in this case, unconfoundedness does, as in all classical randomized designs. We can think to such a design as a set of completely randomized experiments taken within blocks defined by the values of $\boldsymbol{X}$. In section 2.1.2, we noted that (2.29) implies ATT(x) = ATU(x) = ATE(x).

In this case, in order to get an unbiased estimate of the ATE, the (2.26) needs to be adjusted by weighting units with the inverse of the treatment probabilities:

$$\hat{ATE} = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{Y_i^{obs} D_i}{\Pr(D_i = 1 \mid X_i)} - \sum_{i=1}^{N} \frac{Y_i^{obs}(1 - D_i)}{1 - \Pr(D_i = 1 \mid X_i)} \right) \tag{2.30}$$

The estimator in (2.30) is similar to the Horwitz-Thompson estimator, well known in sampling literature. This is an estimator with beneficial asymptotical properties (Hirano et al. 2000).

As we will see in the next section, the last outlined situation is the more similar to an observation studies with a regular assignment.

Before concluding this section we note that experimental designs, as a basis for causal inference, are criticized by some authors. For example, Manski and Garfinkel (1992) note that “experimental evaluation actually requires that a highly specific and suspect structural assumption hold: individuals and organizations must respond in the same way to the experimental version of a program as they would to the actual version". In other words, experiments can be seriously affected by lack of external validity, that is, we are not always allowed to extrapolate results from an experimental setting to a natural occurring one. Anyway, many social research questions cannot be investigated through

experiments, either for ethical or practical reasons (experiments are expensive and time consuming).

2.3 Estimating causal effects in observational studies under regular assignment mechanisms

In the previous section we have seen that in experimental studies the independence hypothesis is plausible when the researcher is able to randomise the assignment of units to control and treatment. In this case, a simple estimator can be used (2.26).

In observational studies, in general, the independence hypothesis is not plausible. In our setting we would be willing to assume that childbearing decisions are purely random or, at least, that characteristics of households deciding to have a child are, on average, equal to those of the other households. The more likely scenario, however, is that the two groups of households will differ quite substantially. If the characteristics that determine the childbearing decision impact also on the consumption growth, which is the outcome of interest, then the simple estimator in (2.26) will give a biased estimate of the childbearing effect. Several methods can be adopted in alternative to handle this problem. Each relies on specific assumptions and requirements that we want to make evident and compare in the remaining of this section.

We organize these methods in two groups. The first one includes methods relying on the assumption of selection on observables, while the methods in the second group address the presence of unobservable confounders. In other words, the first group of methods assume a regular assignment mechanism, whereas the others assume a latent regular one. Methods in the first group allow, in different ways, to make comparisons between treated and untreated units “ceteris paribus”. In other words, the idea is to make comparisons between units in the two groups with the most similar characteristics as possible. These methods include regression, matching, stratification and weighting methods. In the second group we will consider the IV methods.

We start in this section by presenting the first group of methods. The fundamental identifying assumptions in this case are:

$$Y_{i1},\ Y_{i0} \perp D_i \,|\boldsymbol{X_i}\,, \tag{2.31}$$

$$0 < P\,(D_i = 1|\ \boldsymbol{X_i}) < 1. \tag{2.32}$$

Assumption (2.31) is known as unconfoundedness (UNC)[4]. The combination of the two hypotheses (2.31) and (2.32) is referred to as strong ignorability.

The unconfoundedness assumption asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates. In other words, within subpopulations defined by values of the covariates, we have random assignment. This assumption rules out the role of the unobservable variables. The issue of unobserved covariates can be addressed using models for sensitivity analysis (e.g., Rosenbaum and Rubin, 1983b) or using non parametric bounds for treatment effects (Manski, 1990).

Assumption (2.32) implies equality in the support of X in the two groups of treated and controls (i.e. Support($\boldsymbol{X}|D$=1) = Support($\boldsymbol{X}|D$=0)) which guaranties that ATE is well defined (Heckman et al, 1997). Otherwise for some values of the covariates there would be some units in a group for which we could not find any comparable units in the other.

At this point it is interesting to remember the decomposition of the selection bias proposed by Heckman et al (1998). They showed that the selection bias (B) can be decomposed in three components: $B = B_1 + B_2 + B_3$. The first component, B_1, refer to the bias caused by non-overlapping supports of X in the treated and control group. The term B_2 depends on misweighting within the common support, as the empirical distributions of treated and non-treated may not be the same even when restricted to the common support. Finally, the term B_3 is the true econometric selection bias resulting from "selection on unobservables", that is, it is the bias arising from a different distribution of

[4] The unconfoundedness assumption has been referred to also as the conditional independence, selection on observables or the exogeneity assumption (Imbens, 2004).

relevant unobserved variables between treated and controls. Under UNC the term B_3 is zero. The other bias components are cancelled out when we restrict the analysis on the common support (B_1) and we balance covariates in the group of treated and controls (B_2).

It is important to note that identification of ATT requires weaker versions of assumptions (2.31) and (2.32). In particular we need only:

$$Y_{i0} \perp D_i \,|\boldsymbol{X_i}\,, \tag{2.33}$$

$$0 < P\,(D_i = 1 |\, \boldsymbol{X_i}). \tag{2.34}$$

Under unconfoundedness, various alternative estimators have been proposed for the estimation of average causal effects. These estimation methods includes: (i) methods based on estimating the unknown regression functions of the outcome on the covariates (Hahn, 1998; Heckman et al, 1997; Heckman et al, 1998), (ii) matching on covariates (Abadie and Imbens, 2002), (iii) methods based on the propensity score including blocking (Rosenbaum and Rubin, 1984) and weighting (Hirano et al, 2003), (iv) combinations of these approaches, for example weighting and regression (Robins and Rotnizky, 1995) or matching and regression (Abadie and Imbens, 2002), and (v) Bayesian methods, which have found relatively little following since Rubin (1978). All these methods attempt, in different ways, to cancel the bias term B_2. We will discuss in the following, regression and propensity score matching methods.

Clearly, assumption (2.31) may be controversial. It requires that all variables that affect both outcomes and the likelihood of receiving the treatment are observed. Although this is not testable, it can be a very strong assumption in some applications. However, any alternative assumption that does not rely on unconfoundedness while allowing for consistent estimation of the average treatment effects must make alternative untestable assumptions. In the literature, especially econometrical, several approaches are proposed that overcome the UNC hypothesis.

The well-known "Heckit" correction (Heckman selection model; Heckman, 1979) is one of the traditional approaches to dealing with the sample selection problem. It is now widely recognised that without an instrument for selection into the treatment group (in other words, a variable that has explanatory power in a selection equation but does not affect outcomes except through selection) these models are identified only by assumptions about functional form and error distributions. Identification through functional form alone has been shown to be quite tenuous, resulting in standard errors that are often very large, and results that are very sensitive to the particular distributional assumptions invoked (see Puhani, 2000). Alternatively to the methods, semi-parametric and non-parametric IV models have been received an increasing interest.

Another favourable situation for the identification of treatment effects arises when participation into treatment is determined by a Regression Discontinuity Design (RDD). In this design, assignment to treatment solely depends on whether observable pre-intervention variables satisfy a set of conditions known to the analyst. For example, units willing to participate are divided into two groups according to whether or not a pre-intervention measure exceeds a known threshold, but only units scoring above that threshold are assigned to the program. In a neighbourhood of the threshold for selection a RDD presents some features of a pure experiment (see for example, Hahn et al, 2001).

Finally, we mention the Difference-in-Difference estimator (DID), which is well-known in the econometric literature (see e.g., Wooldridge, 2002). DID methods for estimating causal effect of policy interventions are widely used in economics, in particular when outcomes are measured in both the treatment and control group before and after the policy intervention. To apply DID we need that units are observed at least in two time points. Let suppose we have N individuals observed at two time periods t_0 and t_1 and suppose that some units are exposed to the treatment between these two time points. The difference in difference (or "double difference") estimator is defined as the difference in average outcome in the treatment group before and after treatment minus the difference in average outcome in the control group before and after treatment:

$E\left[Y_1^{t_1} - Y_1^{t_0}\right] - E\left[Y_0^{t_1} - Y_0^{t_0}\right]$. An advantage of the DID estimator is that it allows us to control for selection into treatment caused by time-invariant unobserved variables. Heckman et al (1998) propose to combine the DID with propensity score matching methods in order to relax the UNC assumption. In other words, we have not to assume that the bias term B_3 is zero, but we have only to assume that is zero the difference in the bias at the two time points. Therefore, some authors have found the DID-PSM estimator useful arguing that it is more robust since it eliminates temporarily invariant source of bias (Dehejia and Wahba, 1999; Smith and Todd, 2005; Aassve et al, 2007). We use a combination of DID with other methods in the applications we present in chapter 5 and 6, taking advantage on the panel structure of our data.

However, the interpretation of the standard DID estimator depends on the assumptions about the unobserved components. In the traditional DID is supposed that the effect of the time is linear and constant across individuals. Recently, Athey and Imbens (2006) proposed a generalisation of the standard DID method, the change-in-change estimator, in which they allow the effects of both time and intervention differ systematically across individuals.

2.3.1 Regression methods

The assumptions underlying the regression model are well known and outlined in common econometric text books (e.g. Green, 2002; Wooldridge, 2002). However, these assumptions are usually stated out of any framework for causal inference. Then, it is of interest to clarify the regression model under the potential outcomes framework introduced earlier. Let specify a linear model for the two potential outcomes:

$$Y_{i0} = X_i\beta + e_i \tag{2.35}$$

$$Y_{i1} = X_i\beta + \Delta + e_i \tag{2.36}$$

where $\Delta = E(Y_{i1})-E(Y_{i0}) = \text{ATE}$.

The models expressed by (2.35) and (2.36) assume that the relationship between potential outcomes and covariates are linear and that there is no interaction between X and the treatment. In fact, the vectors of parameters in the two regressions are equal. Moreover, the treatment effect is assumed to be constant (in fact Δ is not indexed with i). Substituting the two models for potential outcomes (2.35) and (2.36) in the equation (2.2) we get the traditional linear multiple regression model:

$$\begin{aligned} Y_i^{obs} &= \left(X_i\beta + \Delta + e_i\right) * D_i + \left(X_i\beta + e_i\right) * \left(1 - D_i\right) \\ &= X_i\beta + D_i\Delta + e_i \end{aligned} \tag{2.37}$$

If the true model were non linear, the OLS estimates of the treatment would be in general biased (see the discussion in Goodman and Sianesi, 2005). Moreover if the effect of the treatment changes by unit characteristics (heterogeneous treatment effect) OLS will not in general recover the ATT. In fact, the constant treatment effect assumption implies that ATE coincides with ATT, as previously noted. Both these problems are exacerbated if some units fall outside of the common support of the covariates, that is, if there are units receiving the treatment for which there are no comparable unit receiving the control. In this case, performing OLS might hide the fact that the analyst is actually comparing incomparable units by using the linear extrapolation. Of course, the problem can be circumvented by estimating the common support and running the regression conditioning on it. Moreover, heterogeneous treatment effect can be allowed by including a complete set of interactions between covariates, X, and the treatment, D. This gives rise to the so called Fully Interacted Linear Model[5] (FILM – see Goodman and Sianesi, 2005). We can be obtain the FILM under the potential outcomes framework in this way:

[5] Note that the FILM introduced here is different from a fully saturated model, which is a linear regression of Y on D controlling non parametrically for the full set of main effects and interactions of the covariates X and the treatment D. This is the most flexible form of regression is possible and the most similar to an exact matching on covariates. However, both exact matching and fully saturated regression may not be feasible when the sample is small and/or the set of covariates is large and many of them are multivalued, or, worse, continue. If the number of

$$Y_{0i} = X_i\beta + e_i \tag{2.38}$$

$$Y_{1i} = X_i\beta' + \Delta + e_i \tag{2.39}$$

In the two models (2.38) and (2.39) the parameters of covariates X are allowed to be different. As a consequence we get the model:

$$\begin{aligned} Y_i^{obs} &= \left(X_i\beta' + \Delta + e_i\right) * D_i + \left(X_i\beta + e_i\right) * \left(1 - D_i\right) \\ &= X_i\beta + D_i\Delta + I_{Xi}\phi + e_i \end{aligned} \tag{2.40}$$

where I_X includes all the possible interactions among covariates in X and the treatment indicator, D, while he vector Φ collects the coefficients of variables included in I_X. These coefficients coincide with the difference between the correspondent vectors of coefficients in the model for the two potential outcomes (3.24): $\Phi = \beta' - \beta$. In the model (2.40) we assume that individuals with the same value of X have the same treatment effect, but that the impact of the treatment can differ across individuals with different observable characteristics. In this case, ATE(x) = ATT(x) = ATU(x) but ATE, ATT and ATU are, in general, different. We also note that in this case the parameter Δ does not represent the ATE as in model (2.37) but it represents the effect of D when the variables interacting with it are all equal to zero. The ATE and ATT in this model are, respectively, given by:

$$ATE = \left(\beta' - \beta\right)^T E\left(X\right) + \Delta \tag{2.41}$$

$$ATT = \left(\beta' - \beta\right)^T E_{D=1}\left(X\right) + \Delta \tag{2.42}$$

where T indicates the transpose operator. As we can easily see from the (2.41) and (2.42), under the regression model (2.40) ATE and ATT differ if the effect

---

cells is very large with respect to the size of the sample is possible that some cells contain only treated or only control subjects.

of the treatment interacts with at least one of the covariates included into the model and the mean of this covariate differ in the treated and control groups.

The previous discussion clarifies that we can make the standard multiple regression models increasingly flexible to outrun hypotheses that seem implausible in a given setting. Also the linearity assumption can be relaxed if we use a non-parametric method, such as a kernel estimator (see Hardle and Linton, 1994). Non-parametric methods, however, are affected by some problems (for example inefficiency) when the number set of covariates is large and many of them are multi-valued, or, worse, continue. This problem, known as *curse of dimensionality*, is common also to matching methods that we discuss in the following section.

2.3.2 Propensity score matching methods

Regression is not the only way to deal with selection on observables. Matching estimators are another class of estimators that rely on the same unconfoundedness assumption as regressions. However, the weighting of estimated treatment effects across different individuals remains under the explicit control of the researcher, rather than being implicit in the estimator, as in OLS. Thus, matching methods are likely to be more amenable to heterogeneous treatment effect context.

Matching is an intuitive and appealing method, which basic idea consists of contrasting treated and control units with the same characteristics X. Starting from assumption (2.31), the basic idea is that within each cell defined by the values of the covariate X assignment to treatment or control group is random. Therefore, if in a given application we know, or we are willing to assume, that all relevant variables that impacts the selection on treatment and outcome are collected in the set X (and hence we are confident that assumption (2.31) holds) we can match each treated unit with one (or more) control unit with the same values of X. The group of treated and matched controls will differ only for the exposure to treatment and, therefore, differences in the outcome between the two groups can be attributed to the treatment.

We have already noticed that the fundamental problem of causal inference can be seen as a "missing data" problem. From this viewpoint, matching methods are a way to "impute" missing observations for counterfactual outcomes. Using the missing data terminology, we can say that their validity stands on the assumption that the counterfactual observations are "missing at random" (Rubin, 1974), while randomized experiments ensures that the missing information is "missing completely at random".

Let's re-write the ATE in the following way:

$$\begin{aligned} E\left(Y_1 - Y_0\right) &= E_X\left[E\left(Y_1 - Y_0\right) | X = x\right] \\ &= E_X\left[E\left(Y_1 | X = x\right) - E\left(Y_0 | X = x\right)\right] = \\ &= E_X\left[E\left(Y_1 | D = 1, X = x\right) - E\left(Y_0 | D = 0, X = x\right)\right]. \end{aligned} \tag{2.43}$$

In words, the (2.43) says that the ATE can be calculated as the average, with respect to the distribution of X, of average causal effects calculated in subpopulations defined on the X-values (that is, ATE(x)). However, in order to apply the (2.43) to the group of treated and matched controls we need to have perfectly balanced distributions for X. When the number of matching variables is large and/or when some of X are continuous exact matching becomes unfeasible and a distance metric have to be used to weight comparisons of matched treated and control units. An alternative is to implement the matching on a univariate variable, which "summarizes" the information incorporated in X, as opposed to matching directly on the multivariate set X. Well known are matching methods that use the propensity score, which can be defined as

Definition 2.7 – Propensity score

The propensity score is the conditional probability of receiving a treatment given pre-treatment characteristics:

$$e(X) \equiv Pr\{D = 1|X\} = E\{D|X\}. \tag{2.44}$$

The substitution of the multivariate set X with the univariate $e(X)$ in the matching procedure is justified by the following important theorems due to Rosenbaum and Rubin (1983a):

Theorem 2.1 - Balancing property of the propensity score
Conditioning on the propensity score, X and D are independent:

$$X \perp D \mid e(X).$$

Theorem 2.2 - Ignorability given the propensity score
If treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score; that is

$$(Y_1, Y_0) \perp D \mid X \text{ and } 0 < P(D=1|X) < 1$$

implies

$$(Y_1, Y_0) \perp D \mid e(X) \text{ and } 0 < P(D=1|e(X)) < 1 .$$

Theorem 2.1 states that observations with the same propensity score have the same distribution of characteristics X, independently of treatment status. In other words, for a given propensity score, exposure to treatment is random. When the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates are balanced in expectation across the two groups. Therefore, matching on the propensity score is equivalent of matching on X.

Theorem 2.2 is the key result to show that if treatment assignment is strongly ignorable, then adjusting for $e(X)$ is sufficient to produce unbiased estimates of ATE. On the basis of these two theorems we can substitute in (2.43) $e(X)$ to X:

$$\text{ATE} = E_{e(X)}\left[E\left(Y_1 \mid D=1, e(X)\right) - E\left(Y_0 \mid D=0, e(X)\right)\right]. \qquad (2.45)$$

where the outer expectation is over the distribution of $e(X)$.

In observational studies the propensity score is not known and it has to be estimated from the data available. Since a fully nonparametric estimation of the propensity score would be liable to suffer from the same curse of dimensionality as the standard matching estimator, the estimation task is generally accomplished parametrically. Propensity score matching thus becomes a semi-parametric approach to the evaluation problem

Using the common logit or probit models, we can write $e(X_i) \equiv Pr\{D_i = 1|X_i\} = F[h(Xi)]$, where $F(.)$ is, respectively, the normal or the logistic cumulative distribution and $h(Xi)$ is a function of covariates with linear and higher order terms. The choice of which higher order terms to include, as well as interactions among covariates, is determined solely by the need to balance covariates distribution in the two treatment groups (Dehejia and Wahba, 1999). Simple parametric specifications for the propensity score have indeed often been found to be quite effective in achieving the balancing required (see for example Zhao, 2005).

The estimation of the propensity score is, however, not sufficient to estimate ATE using the (2.45). The reason is that the probability of observing a treated and a control unit with exactly the same value of the propensity score is, in principle, zero, since $e(X)$ it is a continuous variable. Then, we need to use some algorithm to match treated and controls.

Various matching methods have been proposed in the literature to overcome this problem and the most widely used are stratification, nearest neighbour, radius, kernel matching.

The idea of stratification matching is to partition the range of the propensity score into a set of intervals (strata), such that within each interval treated and control units have, on average, the same propensity score. Then, within each interval in which both treated and control units are present, the difference between the average outcomes of the treated and the controls is computed. The ATT, for example, can be, finally, obtained as the average of the

ATT calculated in each block, with weights given by the distribution of treated units across blocks. One of the pitfalls of the stratification method is that it discards observations in blocks where either treated or control units are absent.

The most straightforward alternative matching estimator is the nearest neighbor (NN) matching, which consists of taking each treated (control) unit and searching for the control (treated) unit with the closest propensity score, i.e. the nearest neighbor. Several variants of NN matching are proposed, e.g. NN matching "with replacement" and "without replacement". In the former case, an untreated individual can be used more than once as a match for treated units and *vice versa*, whereas in the latter case each unit is considered only once. Matching with replacement involves a trade-off between bias and variance. If we allow replacement, the average quality of matching will increase and the bias will decrease. Once each unit has found a match in the other group, the difference between the outcomes of the two units is computed. The ATT is then obtained by averaging these differences. Another alternative to the NN matching is the k-NN method, which consists to use more than one ($k > 1$) nearest neighbours. This form of matching involves a trade-off between variance and bias, too. It trades reduced variance, resulting from using more information to construct the counterfactual for each participant, with increased bias that results from on average poorer matches (see e.g. Smith, 2000). Then, the outcome of each unit is contrasted to a weighted average of the outcome of the k-nearest neighbours. This involves another choice concerning the weights to be used.

In the case of the nearest neighbor method all treated units find a match. However, it is obvious that some of these matches are fairly poor because for some treated units the nearest neighbour may have a very different propensity score and nevertheless he would contribute to the estimation of the treatment effect independently of this difference. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper). Bad matches are avoided and the matching quality rises. However, if fewer matches can be performed, the variance of the estimates increases. A variant of caliper matching is the so-called radius matching. The basic idea of this method is to use not only the nearest neighbour within each caliper but all of the comparison members

within the caliper. A drawback of caliper and radius matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

With kernel matching all treated are matched with a weighted average of all controls with weights that are inversely proportional to the distance between the propensity scores of treated and controls. The way the weights are calculated depend on the specific kernel function we use. The widest employed versions are the Epanechnikov and Gaussian kernel.

It is clear from the above considerations that the various methods reach different points on the frontier of the trade-off between quality and quantity of the matches and none of them is *a priori* superior to the others. Asymptotically, all PSM estimators should yield the same results (Smith, 2000), while in small samples the choice of the matching algorithm can be important (Heckman et al, 1997). The performance of different matching estimators varies case-by-case and depends largely on the data structure at hand (Zhao, 2005). Pragmatically, it seems sensible to try a number of approaches. If they give similar results, the choice is irrelevant. Otherwise, further investigation may be needed in order to reveal more about the source of the disparity (Bryson et al, 2002).

We can write a general formula for the matching estimators of ATT, ATU and ATE in the following way. Let denotes with I_0 and I_1 the sets of indices for untreated and treated units, respectively. To estimate the treatment effect for a treated person $i \in I_1$, outcome Y_{i1} is compared to an average of the outcomes Y_{j0} for matched units $j \in I_0$ in the untreated sample. Typically, when the observed propensity score of an untreated person is closer to that of the treated person, using a specific distance measure, the untreated person gets a higher weight in constructing the match. Following Heckman et al. (1997), the estimated gain for unit i in the treated sample is

$$Y_{i1} - \sum_{j \in I_0} W(i,j) Y_{j0} \tag{2.46}$$

where $W(i,j)$ is usually a positive valued weight function, defined so that for each $i \in I_1 \sum_{j \in I_0} W(i,j) = 1$. The choice of a weighting function reflects the choice of a particular distance measure used in the matching method, and the weights are

based on distances in the $e(X)$ space. For example, for each $i \in I_1$ the nearest-neighbor method selects one individual $j \in I_0$ as the match whose $e_j(X)$ is the closest value to $e_i(X)$. On the opposite side, the kernel methods construct matches using all units in the comparison sample and down weighting "distant" observations.

ATT is estimated averaging (2.46) over the sample of treated:

$$\hat{ATT} = \sum_{i \in I_1} w(i) \left[Y_{i1} - \sum_{j \in I_0} W(i,j) Y_{j0} \right] \quad , \tag{2.47}$$

where $w(i)$ are the weights assigned, in calculation of the average, at each unit which satisfy the condition that $\sum_{i \in I1} w(i) = 1$. In the simplest case $w(i) = 1/n_1$ for each $i \in I_1$, where n_1 is the number of treated units in the sample. Different values of $w(i)$ may be used to select different domains of $e(X)$, or in order to account for heteroschedasticity.

Similarly, the ATU is estimated by

$$\hat{ATU} = \sum_{j \in I_0} w(j) \left[Y_{j0} - \sum_{i \in I_1} W(i,j) Y_{i1} \right] \quad , \tag{2.48}$$

where in squared brackets we represented the estimated gain for an untreated unit and $w(j)$ are the weights assigned in the average to effects estimated on the single units.

At this point ATE can be estimated combining (2.47) and (2.48). In fact, writing the ATE as a weighted average of ATT and ATU with weights proportional to n_1 and n_0, it can be estimated by

$$\hat{ATE} = \frac{n_1}{n} \sum_{i \in I_1} w(i) \left[Y_{i1} - \sum_{j \in I_0} W(i,j) Y_{j0} \right] + \frac{n_0}{n} \sum_{j \in I_{01}} w(j) \left[Y_{j0} - \sum_{i \in I_1} W(i,j) Y_{i1} \right], \tag{2.49}$$

where we compare the outcome observed on each treated unit with outcomes observed on some matched untreated and vice versa. When $w(i) = 1/n_1$ for each $i \in I_1$ and $w(j) = 1/ n_0$ for each $j \in I_0$, the (2.49) simplifies to

$$\hat{ATE} = \frac{1}{n}\left\{\sum_{i\in I_1}\left[Y_{i1} - \sum_{j\in I_0} W(i,j)Y_{j0}\right] + \sum_{j\in I_{01}}\left[Y_{j0} - \sum_{i\in I_1} W(i,j)Y_{i1}\right]\right\}. \quad (2.50)$$

Formulas (2.47)-(2.50) will be differently specified according to the specific matching algorithm we choose.

As the estimation of the standard errors of the treatment effects is concerned, it should ideally adjust for the additional sources of variability that are introduced by the estimation of the propensity score, as well as by the matching process itself. For kernel-based matching, analytical asymptotic results have been derived by Heckman et al (1998), whereas for others matching the common solution is to resort to bootstrapped confidence intervals. However, Abadie and Imbens (2004) proved that bootstrap fails with nearest neighbor matching, due to its extreme non-smoothness. The implication for empirical practice of these results is that for methods like nearest neighbor matching one should use analytical variance estimators, such those developed by Abadie and Imbens (2006).

2.3.3 Regression versus propensity score matching methods

We have seen that both regression model and propensity score matching rely on the UNC assumption. However, they are different in the way they estimate causal effects. A fundamental difference between the two methods is that PSM makes more explicit the comparison of treated and control units.

Matching techniques offer a number of practical advantages relative to regression. They are nonparametric and tend to focus attention on the common support condition. However, a principal conceptual difference between regression-based and matching techniques is the flexibility the latter gives the researcher in choosing how to aggregate heterogeneous impacts. In a matching estimator, the weighting is easily manipulated so that interesting parameters (that is, interesting averages) like the average effect of the treatment on the treated,

can be estimated. On the other hand, if treatment effects are homogeneous, then the regression-based estimator is more efficient.

We have seen that investigators using regression based approaches could relax many parametric assumptions, like homogeneous treatment effects. Moreover, it is also possible to incorporate common support concerns in a regression framework. The essential difference between regression and matching remains the weighting scheme used to take the average of the treatment effects at different values of the covariates. Regression gives more weights to cells in which the proportion of treated and non-treated is similar. Matching gives more weights to cells in which the proportion of treated is high when calculating the ATT and, vice versa, it gives more weights to cells in which the proportion of untreated is high when calculating the ATU (Angrist, 1998).

Concluding, one of the most desirable features of the PSM is that it forces the researcher to design the evaluation framework and check the data before looking at the outcomes. They dominate other strategies that require selection on observables, like regressions, because they involve a more convincing and direct comparison between treated and control subjects.

2.4 Estimating causal effects under a latent regular assignment mechanism using Instrumental Variables

So far we have dealt with the problem of self selection where selection in the treatment status depends only on observed covariates. We now move on to the case where we suspect failure of the unconfoundedness assumption. More specifically, we may suspect the presence of some unobserved covariates that influence our outcome and are associated with the selection into treatment. A classic example of such a violation is found in the labour economics literature, whereby the interest lies in estimating the returns to schooling on wages. Here a common problem is introduced by the fact that ability, which is unobserved to the analyst, may influence the outcome of interest, namely individuals' wages. In our case, where the interest lies in estimating the effect of childbearing on

economic wellbeing, we might also encounter unobserved characteristics that will influence the outcome of interest. To better understand why, it is important to bear in mind that consumption expenditures are determined, mainly, by individuals' labour income. As is well known, fertility decisions are, often, endogenous with respect to work decisions, and therefore household income which in turn drives expenditure. This is especially the case for women (Kim and Aassve 2006). Again unobserved ability may enter the picture. On one hand, women's earnings ability will influence their work decisions, which in turn will influence expenditure levels. Unobserved ability, in a general meaning, may also be correlated with contraceptive use and it is likely to influence the outcome of interest.

These situations are known in the econometric literature as *selection on unobservable* and refers to the source of endogeneity driven by *omission of relevant variables*. We briefly start by recalling the concept of endogeneity in econometrics to create a parallel set-up with respect to the potential outcomes framework.

In a regression framework, a given explanatory variable is said to be endogenous if it is correlated with the error term. This correlation can arise from the following causes: (see for e.g. Wooldridge, 2002; pp. 50-51):

- **Relevant omitted variables**: This problem arises when one or more unobserved variables, which have an influence on the outcome, are correlated with one or more covariates. Unobserved variables are included in the error term, giving rise to the correlation between endogenous covariates and the error term.
- **Measurement error**: Let suppose that only an imperfect measure of a given covariate X is available, say X^*. If this is the case, we necessarily put a measurement error into the error term. This situation may or may not give rise to correlation between X and the error, depending on how X and X^* are related.
- **Simultaneity**: It arises when at least one of the explanatory variables is determined simultaneously along with the outcome, Y. If, say X is

determined partly as function of Y, then X and the error are generally correlated.

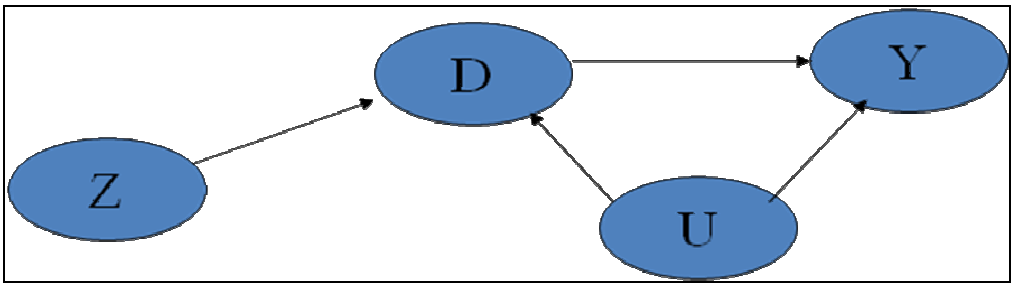
In the sequel we will focus on the first situation, as it appears to be the most serious source of endogeneity in our application. In the potential outcome framework developed before we can think to this situation as the case in which even after conditioning on observed covariates, $\boldsymbol{X}$, there remains a certain degree of dependence among potential outcomes and treatment assignment. Only if we could also condition on unobservables, $\boldsymbol{U}$, will this dependency disappear. In other words, assumption (2.31) has to be relaxed so that:

$$Y_{i1},\ Y_{i0} \perp D_i \,|\boldsymbol{X_i}, \boldsymbol{U_i} \tag{2.51}$$

Instrumental variable analysis is a common method used by econometricians to estimate causal effects in the presence of endogeneity. This method can be used when one or more variables, termed instruments, are available. A valid instrument is a variable that is correlated with the endogenous variable but is uncorrelated with the outcome, given the endogenous variable. In other words, it affects the outcome only through its effect on the endogenous variable. This situation is represented in the figure 2.1, where it is represented a causal link between D and Y. However, this relationship is confounded by the variable U. Since U is unobserved, we cannot use methods like regressions and PSM, which rely on UNC, to estimate the causal effect of D on U. A solution is to use the fact that a variable Z exists so that it influences only D. A change in Z generates an exogenous variation in X, allowing the identification of the causal D-Y relationship.

In many practical applications, however, it becomes hard to find good instruments. Nevertheless, as we will see in chapter 5, we propose two different instruments for the identification of the causal effect of fertility on poverty, without relying on UNC.

Figure 2.1 – A situation where the effect of *D* on *Y* is confounded by an unobservable variable (*U*) and an instrument (*Z*) is available.



In the sequel of the section, we illustrate the identification and estimation of causal effects using IV methods, in a general framework. We consider the simplest case, where only one binary exogenous instrument is available. This is a common situation in many randomised trials when it is not possible to randomise the treatment, but it is possible to randomise the *assignment* to the treatment. In other words, when there is a certain rate of noncompliance.

Noncompliance occurs when the actual treatment that subjects receive differ from their nominal assignment. Here we assume all-or-none noncompliance (as in Angrist et al, 1996): after randomization, some subjects assigned to the treatment will not take it, but effectively take the control, whereas some units assigned to the control receive the treatment. In this situation an instrumental variable is represented by the assignment indicator, *Z*, which is different from the treatment indicator, *D*. This is because the assignment to treatment influences the probability to take the treatment and has no direct impact on the outcome, because of randomisation.

This situation can be viewed also as a randomised encouragement design when the analyst randomises the encouragement to receive the treatment (*Z* = encouragement indicator) instead of the treatment itself (*D*). In this case the instrument is represented by the encouragement assignment (*Z*=1 for units encouraged to take the treatment; *Z*=0 for other units).

In an observational study, like ours, we can use this scheme when the instrumental variable can be considered as exogenous, that is not determined itself by the variables that influence the selection in the treatment and the outcome.

Angrist, Imbens and Rubin (1996) (in the following AIR) use this framework to relate the statistical literature on potential outcomes to the econometrical literature on IV. In this paper the authors make explicit the assumptions needed to identify a causal effect using an IV analysis under relatively weak assumptions. Next section builds heavily on this paper.

2.4.1 Randomized instruments

In this section we treat the case in which the instrument can be thought as exogenous, that is, it is assigned at random and hence it is not confused with the treatment D or the outcome, Y. We indicate with Z_i the assignment received by unit i, which for simplicity will be considered as a binary variable. We indicate with $D_i(\boldsymbol{Z})$ the binary treatment indicator for unit i, which depends on the assignment variable, $\boldsymbol{Z}$. Similarly, the potential outcomes for unit i are indicated as $Y_i(\boldsymbol{Z}, \boldsymbol{D})$. We now give a formal definition of an instrument, both in the econometric and potential outcomes framework.

Let us refer to the *dummy endogenous variable model,* well known in econometrics (see, for example, Maddala, 1983), where a continuous outcome is regressed on a binary endogenous variable, D. Using a latent index formulation, D takes the value 1 if the underlying continuous variable D^*, which is modeled as a linear function of a variable Z, overcome a threshold, conventionally fixed at 0, and D is equal to 0 otherwise. This model can be formulated in the following way:

$$Y_i^{obs} = \beta_0 + \beta_1 D_i + e_i \tag{2.52}$$

$$D_i^* = \alpha_0 + \alpha_1 Z_i + v_i \tag{2.53}$$

$$D_i = \begin{cases} 1 & \text{if } D_i^* > 0 \\ 0 & \text{if } D_i^* \leq 0 \end{cases} \tag{2.54}$$

It is interesting to note that this latent index formulation corresponds to the idea that compliance is a "choice" that decision makers formulate comparing expected utility deriving from the two alternative treatment statuses. In this model the dummy variable D is potentially correlated with the error term, e, implying that it is endogenous, in econometric terminology. In other words, the receiving treatment is not ignorable (Rubin, 1978). In order for Z to be an instrument, econometric theory postulates that it must respect the following conditions:

$$E(Z^*e) = 0 \text{ , } E(Z^*v) = 0 \tag{2.55}$$

$$\text{cov}(D, Z) \neq 0. \tag{2.56}$$

Since the error terms are assumed to be zero mean variables, the first condition (2.55) imposes zero-correlation among Z and the error terms. The requirement that $E(Z^*e) = 0$ and the absence of Z in equation (2.52) implies that any effect of Z on Y must be through its effect on D. Condition (2.56) is equivalent to the requirement that α_1 in the (2.53) is different from 0. These two conditions are well known in econometrics as, respectively, the validity and the relevance of the instrument.

In this simple case, the IV estimator is defined as the ratio of two sample covariances (Durbin, 1954):

$$\beta_1^{IV} = \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)} \quad . \tag{2.57}$$

Under assumptions (2.55) and (2.56), the estimator (2.57) consistently estimates β_1, while OLS is biased and inconsistent (Carneiro et al, 2005). In fact, under these assumptions:

$$\begin{aligned}\text{plim}\,\beta_1^{IV} &= \frac{\text{cov}(Y,Z)}{\text{cov}(D,Z)} = \frac{\text{cov}(\beta_0+\beta_1 D+\varepsilon,Z)}{\text{cov}(D,Z)} \\ &= \frac{\text{cov}(D,Z)\beta_1+\text{cov}(\varepsilon,Z)}{\text{cov}(D,Z)} = \beta_1\end{aligned} \tag{2.58}$$

When both Z and D are binary variables the ratio of the two covariances in (2.57) simplify to:

$$\beta^{IV} = \frac{E[Y \mid Z=1]-E[Y \mid Z=0]}{E[D \mid Z=1]-E[D \mid Z=0]}, \tag{2.59}$$

which is often called the *Wald estimator*, since it first appeared in a paper by Wald (1940) on errors-in-variables problems. The (2.59) can be easily estimated by:

$$\beta_1^{IV} = \frac{\dfrac{\sum_{i:Z=1} Y_i}{\#\{Z=1\}} - \dfrac{\sum_{i:Z=0} Y_i}{\#\{Z=0\}}}{\dfrac{\sum_{i:Z=1} D_i}{\#\{Z=1\}} - \dfrac{\sum_{i:Z=0} D_i}{\#\{Z=0\}}} \tag{2.60}$$

where $\#\{Z=z\}$ indicates the sample dimension of the group with $Z=z$.

The (2.60) is simply the ratio between the difference of the sample averages of Y (numerator) and D (denominator) calculated on subpopulations defined by the Z values.

In econometrics, an instrument that does not respect condition (2.55) is said to be *invalid*. In this case cov(Z, e) is not 0 and from the last term in (2.58) we see that we cannot recover a consistent estimate of β_1. On the other hand, if the correlation between Z and D is low the instrument is said *weak*. In the econometric literature, it is well known the problem of the sensitivity of the IV methods when the instrument is weak (e.g., Klepinger et al., 1995). Intuitively,

we can see from the last term in (2.58) that the problem arise since cov(Z, D), which is found at the denominator, is near zero.

It is important to note that condition (2.58) is untestable. When more than one instrument is available this assumption can be indirectly tested through the so-called overidentifying restriction test (ORT, see Hwang, 1980). Recently, Small (2007) have proposed a sensitivity analysis to assess uncertainty about the validity of instruments when more than one instrument is available. A sensitivity analysis approach with respect to the strength of the instrument is, instead, proposed in Small and Rosenbaum (2007).

In the sequel we adopt the setting used by AIR to give a causal interpretation to the estimator in (2.57). The authors use the potential outcomes framework under which they are allowed to state the identifying assumptions in a transparent manner. In order to do so, AIR substitute the assumptions (2.55) and (2.56), made in terms of disturbances, to the following assumptions cast in term of intrinsically meaningful and potentially observable variables. AIR define an IV as a variable that respect the following conditions:

Assumption AIR.1 - Stable Unit Treatment Value Assumption (SUTVA)
In this context SUTVA requires that the potential outcomes for each unit do not depend on assignment and treatment status for the other units. Formally:

If $Z_i = Z'_i$, then D_i ($\boldsymbol{Z}$) = $D_i(\boldsymbol{Z'})$

If $Z_i = Z'_i$, and $D_i = D'_i$, then $Y_i(\boldsymbol{Z}, \boldsymbol{D}) = \text{Yi}(\boldsymbol{Z'}, \boldsymbol{D'})$.

SUTVA implies that

$Y_i(\boldsymbol{Z}, \boldsymbol{D}) = Y_i(Z_i, D_i)$ and $D_i(\boldsymbol{Z}) = D_i(Z_i)$.

Assumption AIR.2 - Random Assignment

The treatment assignment Z_i is random:

$$Pr(\mathbf{Z} = c) = Pr(\mathbf{Z'} = c')$$

for all c and c' such that $\iota^T c = \iota^T c'$, where ι is the N- dimensional column vector with all elements equal to one.

Assumption AIR.3 - Exclusion Restriction

The treatment assignment Z impacts on Y only through D:

$$\boldsymbol{Y(Z, D) = Y(Z', D)}$$

for all $\boldsymbol{Z}$, $\boldsymbol{Z'}$ and $\boldsymbol{D}$.

By virtue of assumption AIR.3, we can define potential outcomes $\boldsymbol{Y(Z, D)}$ as a function of $\boldsymbol{D}$ alone: $\boldsymbol{Y(Z, D) = Y(D)}$. By assumption AIR.1, we can also write $Y_i(D_i)$ instead of $Y_i(\boldsymbol{D})$. This assumption corresponds to the validity assumption of econometricians.

Assumption AIR.4 - Nonzero Average Causal Effect of *Z* on *D*

The average causal effect of Z on D, $E[D_{1i} - D_{0i}]$ is different from 0. This assumption states in terms of average causal effect what assumption (2.56) states in terms of covariance.

Assumption AIR.5 - Monotonicity

$$D_{1i} \geq D_{0i}$$

for all $i = 1, \ldots, N$.

In order to clarify assumption AIR.5 we need to distinguish four groups of individuals that react in different ways to the assignment to the treatment. The first group is given by individuals that are induced to take the treatment (control) by the assignment to the treatment (control). For these individuals, which are called *compliers*, $D_{1i} - D_{0i} = 1$ and the causal effect of Z on Y is $Y_{1i} - Y_{0i}$. Other individuals do not change the treatment status with the assignment. These individuals can be *always-takers,* if $D_{1i} = D_{0i} = 1$, or *never-takers,* if $D_{1i} = D_{0i} = 0$. In both cases the causal effect of Z on Y is zero by virtue of the exclusion restriction. Finally, individuals that do the opposite of their assignment status are called *defiers*. For these individuals $D_{1i} - D_{0i} = -1$ and the causal effect of Z on Y is $-(Y_{1i} - Y_{0i})$. The four types of units are represented in table 2.1 as the product of the cross tabulation of the observed variables Z and D. As we can see, in each cell we have mixtures of units that we cannot disentangle

Table 2.1 - Type of units by observed variables

Z_i	D_i	
	0	1
0	Never-takers Compliers	Always-takers Defiers
1	Never-takers Defiers	Always-takers Compliers

Monotonicity assumption implies that there are no defiers. This assumption is not sufficient to identify the type of a unit by the observed data. In fact, even if in table 2.1 we drop the defiers, mixtures remains in the diagonal cells. However, it is possible to estimate the proportion of units belonging to the group of

compliers. To this end we need to use the randomization assumption which implies that $D_{0i}, D_{1i} \perp Z_i$ and can be written also as:

$$Pr(\tau_i = t \mid Z_i = 0) = Pr(\tau_i = t \mid Z_i = 1), \tag{2.61}$$

for $t \in \{n, c, a\}, \forall x \in Supp(X)$ and where τ indicates the unit type and n, c and a stand, respectively, for never-takers, compliers and always-takers. The last condition is called by Frolich (2007) unconfounded type, and means that the instrument does not affect the relative size of the subpopulations of never-takers, compliers and always-takers. Therefore the population share of always-takers and never-takers can be written as follows:

$$\tau_a = Pr(D = 1 \mid Z = 0) = E[D \mid Z = 0], \tag{2.62}$$

$$\tau_n = Pr(D = 0 \mid Z = 1) = 1 - E[D \mid Z = 1]. \tag{2.63}$$

Consequently, the proportion of compliers is

$$\tau_c = 1 - (\tau_a + \tau_n) = Pr(D = 1 \mid Z = 1) - Pr(D = 1 \mid Z = 0)$$

$$= E[D \mid Z = 1] - E[D \mid Z = 0]. \tag{2.64}$$

By using (2.61), (2.62) and (2.63) we can estimate the proportions of always-takers, never-takers and compliers even if units belonging to these groups are not identifiable. From (2.63) we can see that the proportion of compliers coincides with the causal effect of Z on D, which can be seen as the difference between the probabilities of taking the treatment for the two groups defined by the Z-values. Moreover, by virtue of assumption AIR.4, the proportion of compliers is different from 0.

Monotonicity is a crucial assumption for identification since, otherwise, the treatment effects for those who shift from nonparticipation to participation

when Z shift from 0 to 1 can be cancelled out by the treatment effect of those who shift from participation to nonparticipation (Imbens and Angrist, 1994)[6].

It is straightforward to note that monotonicity is automatically satisfied in model (2.52)-(2.54) due to the (implicit) imposition of the constant treatment effect assumption. In fact, in this model if $Z_i = 0$, then $D_i^* = \alpha_0 + v_i$ and if $Z_i = 1$, then $D_i^* = \alpha_0 + \alpha_1 + v_i$. Hence, $D_{1i}^* > D_{0i}^*$ since α_1 is positive. In fact, α_1, by assumption is different from 0, and since Z is binary it is always possible to redefine it in order to have a positive sign for α_1, that is in order to make units with $Z_i = 1$ more likely to participate to the program (and this is coherent with the view of Z as the assignment indicator). From $D_{1i}^* > D_{0i}^*$ easily follows, $D_{1i} \geq D_{0i}$.

Monotoniticy assumption is untestable and its validity has to be discussed in the context of a given application. Also exclusion restriction is not informed by the data. On the contrary, we can assess if an instrument is weak. In this setting we have a weak instrument if the proportion of complier (that coincides with the causal effect of Z on D) is low. AIR show that the sensitivity of IV to violations of the exclusion restriction and monotonicity assumptions is stronger as long as the group of compliers is small.

AIR demonstrate that given SUTVA and random assignment assumptions we can obtain unbiased estimators for the average intention-to-treat effects: the average causal effect of Z on Y (that we indicate ITT_{ZY}) and the average causal effect of Z on D (ITT_{ZD}). The ratio between the two intention-to-treat effects coincides with the conventional instrumental variable estimator (2.57):

$$\beta_1^{IV} = \frac{ITT_{ZY}}{ITT_{ZD}}. \qquad (2.65)$$

However, in order to give to (2.65) a causal interpretation we also need the other assumptions, under which AIR demonstrate that

[6] Assumptions alternative to monotonicity can be stated. For example, traditional IV methods rely on the constant treatment effect assumption or the assumption that for a given value of Z the probability of participation is 0.

$$\beta_1^{IV} = \frac{ITT_{ZY}}{ITT_{ZD}} = E[Y_{1i} - Y_{0i} \mid D_{1i} - D_{0i} = 1] \tag{2.66}$$

By using equation (2.65) we interpret the IV estimand as the average causal effect calculated on the sub-population of units with $D_{1i} - D_{0i} = 1$, that is for compliers. The effect is what Angrist and Imbens (1994) call the Local Average Treatment Effect (LATE) and is referred to also as CATE (compliers average causal effect). This is, in general, different from the ATE. A serious drawback of the IV estimand given by (2.66) is that it refers to a subpopulation (compliers) that is not identifiable by the data. The interest of the researcher could be, on the contrary, in the estimation of the ATE (or ATT) but this parameter is not identifiable, unless, as noticed by Moffit (1996), we are willing to make additional strong assumptions. For example, Angrist and Imbens (1991) assume that $D_{0i} = 0$ for each unit. This happens in some clinical randomized trials where patients are allocated into two groups by a random mechanism. Patients in the first group receive standard treatment (control); those in the second group are asked if they will accept the experimental therapy (treatment under study). If they decline (do not comply), they receive the standard treatment. Similarly, when $D_{1i} = 1$ for everyone, LATE coincides with the ATU. If these two conditions are simultaneously satisfied, implying that there are only compliers in the population (perfect compliance), LATE banally coincides with ATE. Another situation in which ATE coincide with LATE is in the case of homogeneous treatment effects.

We have to note, moreover, that the effect identified by the (2.65) depends on the instrument we used. So, if several instruments are available the LATE calculated using one instrument is, in presence of heterogeneous effects, different from LATE calculated with another one. This is because each instrument identifies a different sub-population of compliers.

For the previous reasons LATE has been criticized to be a parameter of questionable policy value (Heckman, 1996). However, AIR makes clear that LATE is the parameter that can be identified for the largest subpopulation for

which data are directly informative. In the instrumental variable setting, extensions to groups other than compliers can only be by extrapolation.

In some particular applications the LATE can be a parameter of specific policy interest. This is the case if the policy we want to evaluate coincide exactly with the instrumental variable we use (Carneiro et al, 2005). We provide an example in chapter 5.

Before going ahead with the discussion of IV methods with covariates, we notice that recently Frangakis and Rubin (2002) introduced the principal stratification approach, which is a general framework for comparing treatments adjusting for post-treatment variables. Principal stratification with respect to a post-treatment variable is a cross-classification of subjects defined by the joint potential values of that post-treatment variable under each of the treatments being compared. Principal effects are causal effects within a principal stratum. The key property of principal strata is that they are not affected by treatment assignment, and therefore, can be used just as any pre-treatment covariate. In so far, the principal stratification approach overcomes the problems inside the traditional approach to adjust for post-treatment variable. Frangakis and Rubin (2002) argue that this approach allows to address possible complications in a study, such as the censoring by death, the presence of missing outcomes and treatment noncompliance. In this last respect AIR's approach to adjusting for noncompliance is a special case of the principal stratification framework, where the compliers are a stratum in the principal stratification with respect to the post-treatment variable "compliance behaviour".

2.4.2 Conditionally randomized instruments

In the previous section we relied on the hypothesis that the assignment to treatment, Z, was randomized or, in other words, the instrument is exogenous. However, in many applications Z itself is confounded with D and/or Y. For example, parental education is often used as an instrument in labor economics to identify the returns to schooling. But this variable is likely to be correlated with parent's profession, family income and wealth, which may directly affect the

wage of their offspring. The ability to control for covariates in such situations is important because instruments require conditioning on a set of covariates to be valid.

The conventional approach to accommodate covariates X in IV estimation consists of parametric or semi-parametric methods (2SLS is the most commonly used method of estimation), relying on delicate functional form assumptions (e.g. Angrist and Imbens, 1995; Hirano et al, 2000).

Recently, completely nonparametric identification and estimation in IV models has received a lot of interest. Frolich (2007) pointed out that many nonparametric IV models including covariates also rely on strong assumptions, which could be untenable in most applications. For example, several proposal use additive separability in the error term, that amounts to assume a constant treatment effect, which is also one of the most dangerous assumption of the traditional 2SLS (see e.g., Newey et al, 1999; Das, 2005). In non-separable models identification often require that the instrument is sufficiently powerful to move the value of D_i for every unit i over the entire support of D (e.g., Blundell and Powell, 2003; Florens et al 2002). However, it is hard to assume, in many applications, that such an instrument exists. On the contrary, only a sub-population reacts to the instruments and, as pointed out in the previous section, only on this sub-population we can identify an average causal effect (LATE).

In order to identify non parametrically the LATE accommodating for covariates we can think, intuitively, to impose the assumptions used by AIR in cells defined by the covariates. This is, basically, the approach taken by Abadie (2003) and Frolich (2007). In particular, Frolich (2007) suggests a conditional LATE estimator, which is a completely nonparametric procedure. Another important aspect is that it does not rely on separability. Therefore, this estimator overcomes the pitfall of the previous approaches that we have remembered before.

On the basis of some assumptions, which are basically the conditional version of AIR's assumptions, Frolich identifies the conditional LATE as:

$$LATE\,(x) = E[Y_1 - Y_0 \mid X = x, \tau = c]$$
$$= \frac{E[Y \mid X = x, Z = 1] - E[Y \mid X = x, Z = 0]}{E[D \mid X = x, Z = 1] - E[D \mid X = x, Z = 0]} \quad . \qquad (2.67)$$

Then, the author demonstrates that the marginal LATE can be calculated as follows[7]:

$$LATE = E[Y_1 - Y_0 \mid \tau = c]$$
$$= \frac{\int \big(E[Y \mid X = x, Z = 1] - E[Y \mid X = x, Z = 0]\big) f_x(x)dx}{\int \big(E[D \mid X = x, Z = 1] - E[D \mid X = x, Z = 0]\big) f_x(x)dx} \quad . \qquad (2.68)$$

We note that formula (2.67) is similar to (2.59) but with both numerator and denominator conditional on *X*, confirming the intuition that the conditional LATE can be obtained similarly to the marginal version calculated in cells defined by the *X*-values. In principle, any nonparametric technique could be used to estimate the (2.68). However, when the number of covariates included in the set *X* is high, or include continuous variables, nonparametric estimation becomes difficult, especially in small samples. An alternative, in these situations, is to make use of the balancing property of the propensity score that allows us to substitute the high dimensional set *X* in (2.68) by a univariate variable. In fact, since adjusting for the distribution of *X* is equivalent to adjusting for the distribution of the propensity score, $\pi(x) = P(Z=1|X=x)$, we can write

$$LATE = \frac{\int \big(E[Y \mid X = x, Z = 1] - E[Y \mid X = x, Z = 0]\big) f_x(x)dx}{\int \big(E[D \mid X = x, Z = 1] - E[D \mid X = x, Z = 0]\big) f_x(x)dx}$$
$$= \frac{\int \big(E[Y \mid \pi(X) = \pi, Z = 1] - E[Y \mid \pi(X) = \pi, Z = 0]\big) f_{\pi(x)}(\pi)d\pi}{\int \big(E[D \mid \pi(X) = \pi, Z = 1] - E[D \mid \pi(X) = \pi, Z = 0]\big) f_{\pi(x)}(\pi)d\pi} \qquad (2.69)$$

[7] It is important to note that a common support assumption is needed as stated by Frolich: Supp(*X*/*Z*=1) = Supp(*X*/*Z*=0).

where $f_{\pi(x)}$ is the density function of $\pi(x)$ in the population.

An estimator of (2.69) can be obtained as the ratio of two propensity score matching estimators, measuring, respectively the two intentions to treat effects of Z on Y (numerator) and of Z on D (denominator). Hence, to estimate the numerator (ITT_{ZY}) we consider the variable Z as it would be the treatment and Y as the outcome. For the denominator (ITT_{ZD}) Z is still considered the "treatment" and D the outcome:

$$\hat{LATE} = \frac{\hat{ITT}_{ZY}}{\hat{ITT}_{ZD}} \quad . \tag{2.70}$$

Obviously, the difference between (2.70) and (2.64) is that here we are controlling for covariates, both in ITT_{ZY} and ITT_{ZD}.

Chapter 3
Causal inference in a multilevel setting

Introduction

In this chapter we investigate the methodological modifications needed to the potential outcome framework developed in chapter 2 when we are interested in the estimation of causal effects in a multilevel setting. The chapter is organized as follows. Section 3.1 introduces the basic terminology and the motivations underlying the need for particular care when our data set shows a multilevel structure. Section 3.2 characterizes the most important aspects to be faced when causal inference is implemented in a multilevel setting. Section 3.3 briefly introduces the linear multilevel model. Section 3.4 originally analyses multilevel models in the light of the potential outcomes framework with the goal of emphasize some pitfalls of these models in recovering causal effects. Finally, section 3.5 extends the potential outcomes framework in a multilevel setting keeping into account the critical aspects introduced in section 3.2.

3.1 Motivating multilevel reasoning and multilevel analysis

In many fields, including the social, medical and biological sciences, multilevel structured populations are the norm. Typically these structures are naturally occurring ones. Education provides a prototype example. Pupils or students are grouped in classes; classes are nested within schools; and schools may be administered within local authorities or school boards. The units in such a system

lie at four different levels of a hierarchy. Pupils are assigned to level 1, classes to level 2, schools to level 3 and authorities or boards to level 4. Units at one level are recognized as being grouped or nested within units at higher levels. Such hierarchies are often described in terms of clusters of level 1 units within each level 2 unit etc, and the term *clustered population* is used. In a two-level data structure, to indicate the units at the first level, the terms elementary, level-1 or micro-level units are employed. The first level is also termed *micro level*. The units at the second level are referred to as secondary units, level-2 units, macro-level units or clusters. The second level is also termed *macro level.*

Other examples of hierarchical populations are people within households, within areas; animals within herds, within farms. In other cases the multilevel structure may result from research designs, as in multi-centre clinical trials where patients are nested within clinics. In yet other cases, the data may not obviously seem to be nested, but viewing it as such may yield new insights or more efficient analysis techniques. Examples are repeated measures designs, where measurements are "nested" within individual subjects and multivariate response data where measurements are "nested" within individuals.

Why multilevel structures have captured the interest of many authors in different disciplines? The motivations of what we can label as *multilevel reasoning* are several. In a multilevel structure it is of interest to analyse the interrelationship existing between the different levels and take into account the variability associated with each level of the nesting. To be more concrete, if we refer to the framework of the social analyses, it is often interesting to explore the impact of the *context* on individual level outcomes[8]. The notion of context is quite general and can include spatial contexts (e.g. countries, states, and communities), temporal contexts (i.e. history), organizational contexts (classrooms, schools, firms) and socio-economic contexts (ethnic groups, social classes, economic sectors). The idea that individual responds to their context is a defining claim of the sociological discipline, which dates back at least to Marx's work on political economy (Marx, 1846), to Durkheim's studies of the impact of

[8] The study of the impact of micro level variables on macro level variable can be appropriately studied only obtaining an aggregate measure of the micro-level variable. Otherwise, we would force the macro variable, which varies only at the macro level to vary also at the micro level.

community characteristics on suicide (Durkheim, 1897) and to Weber's research on how religious communities shape economic behaviour (Weber, 1905). However, the relationship between individuals and the society is not unidirectional. In fact, the interactions among individuals and the groups to which they belong consist not only of impacts that groups have on individuals but also of influences that individuals make on groups.

More recently the influence of groups on individual outcomes has been investigated in many fields. For example, Roberts (1999), Macintyre (2000) and Subramanian et al (2001) investigated the impact of the socio-economic context on health. In Demography, Casterline (1985) studied the effect of community characteristics on the reproductive behavior of women in developing countries; Entwisle et al. (1989) analyzed the role of village context in the use of contraceptives; Hirschman and Guest (1990), investigated the impact of contextual variables on fertility in Asia. Many poverty analyses considered, in very different ways, the role of the context on the individual or household living standards. Among them, Van de Walle (1996), Glewwe et al (2002) and Ali and Pernia (2003) stressed the importance of infrastructures (macro level characteristics) in the process of poverty reduction. Mukherjee and Benson (2003) and Justino and Litchfield (2004), studied both individual and community determinants of poverty[9].

All the previous studies recognise the importance of hierarchical structures and have as a common goal to study the interactions that exist between the different levels of these. We refer to such type of analyses as *multilevel research.* In this kind of research, variables can be defined at each level of the hierarchy. For example, in a two level structure, consisting of household clustered in communities, we can have variables measured at household and community level. A detailed characterisation of the different types of variables is outside the scope of this work. We refer instead to the discussion in Hox (1995). However, we find useful to distinguish between two types of macro variables when we analyse data at the individual level. The first type of variables, which

[9] To the best of my knowledge, up to now poverty analyses have not used multilevel modelling techniques.

are usually termed *contextual variables*, consists of variables that represent a feature of the macro unit (cluster) with no corresponding micro level measure. The presence of a road, or other infrastructures, in a community is an example of such a variable. These variables are included into the analysis at the individual level by a process that is called *disaggregation*, because data on higher level units are disaggregated into data on lower level units. The resulting variable is called contextual, because it refers to the higher level context of the units we are investigating. Other variables, named *compositional*, are macro indicators obtained through aggregation of micro level measures. An example is the average size of households residing in the community. These variables are constructed by *aggregation* of data on lower level units into data on higher level units. The term compositional refers to the fact that these variables are a summary measure of characteristics of lower level units that compose clusters.

The previous distinction is linked to the common observation found in multilevel socio-economic research on the importance to distinguish between two sources of variation in the outcome at cluster level: contextual (relating to differences in specific areas' characteristics) and compositional (relating to characteristics of the households or individuals living in different places). That is, if we ask why clusters are different, the answer has to be considered both in their proper features and in the characteristics of units that belong to them. In the context of our application we could ask: what explains the geographic variation in poverty rates? Is it composition - clustering of households with high poverty propensities in certain geographic locations, for instance? Or is it context - something about the socio-economic or institutional setting in those areas - that yields worst wellbeing in some areas than others? A question more closely related to the aim of the thesis would be: what explains the geographic variation in the impact of fertility on poverty? Is it composition - clustering of households with high negative treatment effect in certain geographic locations? Or is it context - something about the socioeconomic or institutional setting in those areas - that makes more difficult to face childbearing events in some areas than others?

For the purpose of applying multilevel models, it is usually not necessary to have well clear in mind at which level each variable is measured. This is important from a conceptual point of view. Historically, multilevel problems have led to approaches that move all variables to one level by disaggregation or aggregation, following by a standard analysis method like an ordinary multiple regression. These kinds of analyses imply statistical and conceptual problems. When we aggregate the micro-level data to the macro-level (for example, we run a regression of the cluster mean of the dependent variables on the cluster means of the covariates) the result is that different data values for many level-1 units are combined into fewer values for level-2 units. As a result, information is lost and statistical analyses loose power. Another statistical problem is that the reliability of an aggregate variable depends, among others, on the number of micro-level units on which is calculated, and thus will be larger for the larger clusters than for the smaller ones. Aggregating analysis can lead also to conceptual errors. The first potential error is the "shift of meaning problem": a variable that is aggregated to the macro level refers properly to the macro-units, not to the micro-units. These variables cannot be used to investigate micro-level relationships. A second potential error is the well known "ecological fallacy" (Robinson, 1950). It consists of failing to recognize that a relationship found between variables at macro level cannot be used to draw conclusions at the micro level. Another problem with aggregated data is that we cannot study potentially interesting cross-level interaction effects between a macro and a micro variable.

On the other hand, if data are disaggregated and variables measured at the macro level are present, we have as a consequence what Snijders and Bosker (1999) call the "miraculous multiplication of the number of units". This refers to the fact that the correct sample size for the cluster level variables is the number of clusters and not the total number of units in the sample, which results if we use a mere disaggregated analysis. The implication is, of course, that significance tests reject the null hypothesis more often than the nominal alpha level indicates. In other words, we can find more spurious significant effect for the macro level variables. Another important statistical problem is in action even if no macro variable is included into the analysis. We have to note, in fact, that

units belonging to the same cluster are likely to be more similar than units belonging to different clusters. This is in line with the presumption that clusters have, in a broad sense, an effect on micro units. For example, in social contexts, people living in the same community share the same socio-economic, cultural and institutional environment. Therefore, there is dependency among units within communities; this violates the independence hypothesis made in standard statistical models. As a result, standard errors will be biased. In particular, they are often underestimated leading, again, to possible spurious effects.

Finally, a potential error similar to the ecological fallacy can be made. This is termed "atomistic fallacy" and consists of drawing conclusions based on macro level relationship from associations found on micro level variables.

Some of the previous issues can be addressed by using care or correction methods. For example, the violation of the independence assumption can be addressed by using certain methods for correcting the standard errors, such as the Huber-White robust standard errors estimation (Wooldridge, 2002).

However, when the multilevel structure is not a mere nuisance factor but a feature of the population of research interest, we need adequate modeling techniques, such as multilevel models. These models allow to bring together, simultaneously, variables at different levels of a hierarchical structure in the same model. Multilevel analysis allows the simultaneous examination of the effects of macro level and micro level variables on micro level outcomes while accounting for the non-independence of observations within groups. Multilevel analysis also allows the examination of both between group and within group variability as well as how group level and individual level variables are related to variability at both levels.

One of the main attractions of multilevel models, as we will see in the next section, is that they allow regression coefficients to vary by clusters. This is obtained by modeling these coefficients as random variables, whose variances has to be estimated, indicating to what extent the relationship varies by groups. An alternative method is given by fixed effects models where the coefficients for each group are treated as fixed unknown parameters to be estimated. Fixed effects models have several disadvantages. First of all, when groups are many,

fixed effect models are not parsimonious, because they involve the estimation of one parameter for each cluster (minus one). By using such models, it becomes unfeasible to include group level variables, because they would be redundant given the fixed group effect. From a conceptual point of view, we can say that fixed effects already explain all the between-group variability and, therefore, there are no residual differences between the groups that can be explained by group level variables. In other words, the between first level unit model is saturated.

Another crucial point against fixed effects models is that they are conceptually meaningless, if the clusters we have in the data are a sample from a larger population of clusters and researcher whishes to draw conclusions pertaining this population. In fixed effects models groups are regarded as unique entities. In multilevel analyses, groups or contexts are not treated as unrelated but are conceived as coming from a larger population of groups about which inferences want to be made. The use of fixed effect models is, on the contrary, correct if we have surveyed data on all the macro level units or we are only interested in the clusters we observed. However, multilevel models treating group-varying coefficients as random variables have the disadvantage of imposing that they follow a probability distribution, usually assumed normal. Normality assumption is not dangerous if the number of clusters is not small (Maas and Hox, 2004).

From the previous discussion we can get many motivations for using a multilevel approach in our work. In the context of our application, that is the study of the causal impact of fertility on poverty using data from the VLSMS, we clearly have a multilevel data structure: household are clustered into communities and communities are grouped into regions. Another level of clustering concerns the waves: we have two measures nested within each household. The multilevel dimension of the data we use was highlighted in section 1.5 where we have described the VLSMS. The fundamental structure we handle consists of two levels: households within communities. The clusters (communities) we observe are a sample drawn from the population of Vietnamese communities (this is literally true because of the two-stages

sampling design of the VLSMS which at the first stage extract a sample of communities). However, we are not specifically interested in the communities surveyed but we wish to do inference about the whole population of communities and hence a random effects approach has to be preferred to a fixed effects one. Moreover, we have important community level variables that we want to include in our analyses.

Also conceptual reasons motivate our choice. Household residing in the same community share the same infrastructures (roads, markets, hospitals, etc.), the same institutions, and the same cultural and physical environment. In so far, there is within-community dependency that we want to keep into account to avoid problems that we discussed before.

Communities can differ in the overall level of poverty and fertility (justifying, as we will see better in the next section, the inclusion of a random intercept). More interestingly, the main relationship of interest in our work, that is the effect of fertility on poverty, could vary by community (requiring the inclusion of a random slope). In other words, it could there be some community "effect", in a broad sense, which can lead the impact of childbearing events on poverty to be different according to the place the household resides.

3.2 Why keeping into account the multilevel dimension in the estimation of causal effects?

Let suppose to have a two-level population. For example, in our case we have households (first level units) grouped in communities (second level units or clusters). If we are interested in the estimation of causal effects in such a population it could be important, from a methodological point of view, and interesting, from a substantive point of view, to keep explicitly into account this multilevel data structure. The motivations can be characterized, in general terms, as follows:

1. Cluster-heterogeneity of the treatment effect,

2. The multilevel nature of the selection process,
3. Potential violation of the SUTVA.

The first aspect refers to the fact that the effect of the treatment can vary substantially according to the cluster to which units belong. In the context of our application the causal effect of childbearing events on poverty can vary considerably according to the community where households live. This can be due to interactions among childbearing events and certain community characteristics (observed or unobserved).

The second issue is that the selection process itself can be multilevel, in the sense that it depends on (observed and/or unobserved) characteristics of clusters that could impact on the cluster average probability of being treated and/or on the effect that same covariates have on this probability. The statistical implication is that controlling (or balancing in the context of matching methods) only for the observed covariates could not be sufficient.

The third issue relates to the potential invalidity of the SUTVA in a multilevel setting. The reasons that make us suspect about the validity this assumption, in a multilevel setting, can be several and depend on the specific studied context and phenomenon. In general, we can suspect that sharing of and competition for resources generates interference among units belonging to the same cluster, while the interference among units in different clusters is absent or negligible.

These three aspects are often confused or at least not distinguished in the literature but from a conceptual point of view it is important to do so. Each of these implies, in fact, different methodological challenges and specific substantive points of interest. These aspects will be minutely discussed and developed in section 3.5. However, we found useful introducing them at this point in order to keep in mind in the following discussion what are the main motivations that push us to explicitly consider the multilevel dimension in a causal inference study. In the next section, we introduce the traditional multilevel model without explicitly consider any issue about causal inference. Then, in section 3.4 we will ask if multilevel models are suitable to answer the specific

methodological and substantive concerns that causal inference poses in a multilevel setting. We will show that the three issues we have introduced in this section are not fully faced by multilevel models. In section 3.5 we propose some combinations of multilevel models and matching methods to better analyse causal effects in a multilevel setting.

3.3 The traditional multilevel linear model

In this section we present a brief overview of the linear multilevel modelling building on the affirmed literature existing on the subject (for example Goldstein, 1995; Hox, 1995; Snijders and Bosker, 1999; Skrondal, and Rabe-Hesketh, 2004). This is also helpful to introduce the notation we will use in the following.

Multilevel models are used to appropriately analyse clustered data structures we introduced in the section 3.1. Here, to simplify the discussion, we consider a two-levels data structure in which N micro units at the first level, indexed by i (i = 1, 2, ... , n_j), are nested in J macro units at the second level, indexed by j (j = 1, 2, ..., J). We allow for an unbalanced data structure. That is, the number of elementary units belonging to a cluster (n_j) is not fixed but change with the cluster.

In order to see the benefits of using a multilevel model, let start specifying a standard linear regression model considering for simplicity one single covariate:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} \tag{3.1}$$

One of the assumptions underlying the regression model in (3.1) is the independence among the observations. The errors ε_{ij} are assumed identically and independently distributed (*iid*). In particular, they are assumed to be normal with zero mean and a standard deviation equal for each observation (omoschedasticity). But as we discussed in the previous section, if observations

are clustered the independence condition is violated. Multilevel models face this within-cluster dependency by decomposing the error term into two additive components, each defined at one of the two levels:

$$\varepsilon_{ij} = e_{ij} + u_{0j}$$

where e_{ij} represents level-1 residual terms and differs between units and clusters, u_{0j} represents the level-2 residual and varies between clusters but is the same for all the units belonging to the same cluster. The second level error, models the combined effect of omitted second level variables, in other words, the unobserved heterogeneity at that level. It can be viewed as a latent variable shared by all units belonging to the same cluster inducing positive intra-class correlation. The reason motivating the subscript 0 for u_{0j} will became clear soon. Our regression model now looks like:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \left(e_{ij} + u_{0j}\right) \tag{3.2}$$

that could be rearranged as:

$$Y_{ij} = \left(\beta_0 + u_{0j}\right) + \beta_1 X_{ij} + e_{ij} \tag{3.3}$$

that makes clear that adding a cluster specific random error allow us to have a random intercept. In model (3.3) β_0 represents the average intercept, while u_{0j} represents the deviation of cluster j's intercept from the mean intercept. It is important to note that, the cluster effect u_{0j} is not a parameter but the realisation of random variable usually assumed normally distributed with zero mean and variance Ψ_{00} which is estimated together with the variance of the first level error term, which we indicate with θ. The fundamental assumptions are resumed as follows:

$$e_{ij} \mid X_{ij} \sim N(0, \theta)$$

$$
\begin{aligned}
Cov\left(e_{ij}, e_{i'j}\right) &= 0, \quad i \neq i' \\
u_{0j} \mid X_{ij} &\sim N\left(0, \Psi_{00}\right) \\
Cov\left(u_{0j}, e_{i'j}\right) &= 0 \\
Cov\left(u_{0j}, u_{0j'}\right) &= 0, \quad j \neq j'
\end{aligned}
\tag{3.4}
$$

The first and third assumptions imply, respectively, that the error terms at the first and second level are uncorrelated with the covariate. The second assumption corresponds to the usual hypothesis of independence among individual error terms while the fifth implies that clusters are independent[10]. The fourth imposes no correlation between the errors terms defined at the different levels.

The variance of the error component is now decomposed into two additive components: the between cluster variance (Ψ_{00}) and the within cluster variance (θ)

$$
\text{var}\left(\varepsilon_{ij}\right) = \text{var}\left(e_{ij}\right) + \text{var}\left(u_{0j}\right) = \theta + \Psi_{00}
$$

A summary measure of the importance of clusters is the proportion of the between cluster variance out of the total variance:

$$
\tau = \frac{\Psi_{00}}{\Psi_{00} + \theta} \tag{3.5}
$$

which is called the *variance partition coefficient* (VPC). In this simple case of a two-level random intercept model, the VPC is equal to the *intra-class correlation coefficient* (ICC) which is usually found with the use of the symbol ρ. This coefficient measures the correlation between two first level units belonging to the same cluster and can be viewed as an indicator of the "closeness" of observations belonging to the same cluster relative to the

[10] Note that having assumed normality, uncorrelation implies independence.

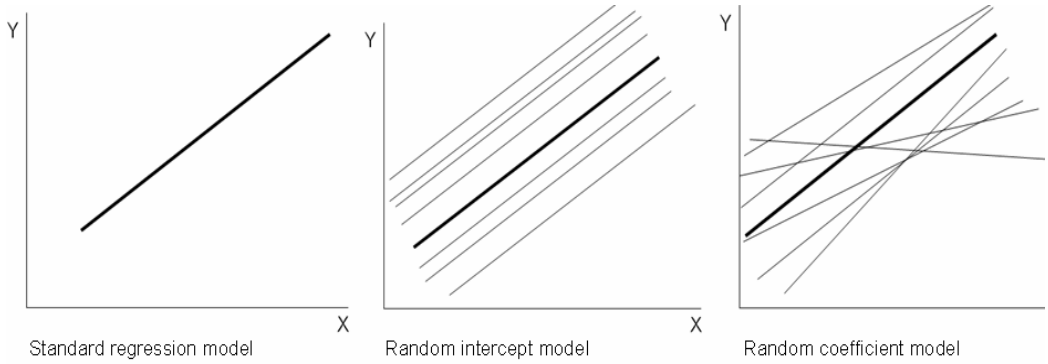
"closeness" of observations belonging to different clusters. The higher is ρ the most important is the clusterisation. It is important to note that in models including covariates as (3.3), ρ is better termed *residual intra-class correlation coefficient* because it represents the correlation between the *Y*-values of two randomly drawn units belonging to the same cluster after controlling for covariates[11].

In the model (3.3) we have parallel regression lines for the different clusters. Hence, the underlying hypothesis is that in all clusters the effect of the covariate *X* on the outcome *Y* is the same. But this may not be realistic in some applications and moreover it would be an interesting to study the interaction between cluster and the *X* effect. In other words researchers could be interested in testing if the influence of *X* on *Y* changes by cluster. Figure 3.1 gives a visual comparison of three ways of modelling the relationship between *X* and *Y*: a simple regression model with a unique (fixed) intercept and a unique (fixed) slope supposed valid for all units in all clusters; a random intercept model with a unique (fixed) slope supposed valid for all clusters while the intercept is allowed to vary around the average represented by a bold line; a random coefficients model, in general the most realistic, that allows both the intercept and the slope to vary cluster by cluster.

It is important to stress the fact that with a multilevel model we do not estimate separate regressions in each cluster. Multilevel models allow us to estimate cluster-specific coefficient regressions but at the same time they allow us to recognise the existence of common factors that make clusters linked each other and make inadequate an analysis that consider clusters independently. In multilevel models, each cluster has each proper regression but these are linked to each other by the fact that they come from a common super-distribution. This amount to assume that the groups we have in the data are a sample of a population of groups.

[11] Intra-class correlation coefficients can be defined also for more complicated data structures involving more than two levels. However, in those cases we can define several types of intra-class correlation coefficients.

Figure 3.1 – A graphical comparison between 3 three ways of modelling a relationship.



The approach that estimates separate regressions for each cluster, furthermore, suffers of another drawback: if in a clusters there are few observations then the sample size for the estimation of the regression in that cluster is small. Multilevel models, instead, can allow data structures containing even one single unit in some clusters.

In multilevel models the estimates for cluster with small size are consolidated by taking information by the other groups, through the mechanism known as *"borrowing strength"* (Kreft and De Leeuw, 1998). Let see from an algebraic point of view how our model changes when we want to allow for an interaction between the effect of the covariate X_{ij} and the cluster effect. This is obtained by including in the model (3.3) an interaction between X_{ij} and a latent variable at the cluster level:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \left(e_{ij} + u_{0j} + u_{1j} X_{ij}\right) \quad (3.6a)$$

$$= \left(\beta_0 + u_{0j}\right) + \left(\beta_1 + u_{1j}\right) X_{ij} + e_{ij} \quad (3.6b)$$

In the first equation (3.6a), it is emphasized how the error component has became more complex and in the rearranged second equation (3.6b), it is made evident that adding the interaction between the random variable u_{1j} and X is equivalent to specify a random slope for this covariate. Likewise u_{0j} also u_{1j} is assumed to be a random normal variable with zero mean and variance Ψ_{11} to be

estimated. The new model is based on the previous assumptions (3.4) plus the following

$$\begin{aligned} u_{1j} \mid X_{ij} &\sim N\left(0, \Psi_{qq}\right) \\ Cov\left(u_{1j}, e_{i'j}\right) &= 0 \\ Cov\left(u_{1j}, u_{1j'}\right) &= 0, \quad j \neq j' \end{aligned} \tag{3.7}$$

The variance-covariance structure at the second level could be specified in several ways. It is usual to allow for correlated random effects defined at the same level. Hence, the variance-covariance matrix for the random effects at the second level could be written as

$$\mathbf{\Psi} \equiv COV\left(u_{0j}, u_{1j}\right) = \begin{bmatrix} \Psi_{00} & \Psi_{01} \\ & \Psi_{11} \end{bmatrix}$$

where Ψ_{00} and Ψ_{11} represent the variances of the two random variables, respectively u_{0j} and u_{1j}, and Ψ_{01} represents the covariance between them.

A very useful way to represent multilevel models is the so-called *multi-stage specification*. This consists in specifying a model for the first level (micro model) and separate equations for each random effect at higher levels (macro models):

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \end{aligned}$$

If we replace the two second level models into the first level model we obtain the so-called *reduced form*, which in this case looks exactly as the equation (3.6b).

An interesting feature of multilevel models is that we can "explain" a random effect at higher levels by including higher level variables. In our case,

we can regress the random intercept and the random slope on a cluster covariate C_j:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$$
$$\beta_{0j} = \beta_{0\_0} + \beta_{0\_1} C_j + u_{0j}$$
$$\beta_{1j} = \beta_{1\_0} + \beta_{1\_1} C_j + u_{1j}$$

The reduced form will be:

$$Y_{ij} = \left(\beta_{0\_0} + \beta_{1\_0} X_{ij} + \beta_{0\_1} C_j + \beta_{1\_1} X_{ij} C_j\right) + \left(e_{ij} + u_{0j} + u_{1j} X_{ij}\right) \qquad (3.8)$$

It is easy to see that adding a cluster level covariate in the random intercept model is equivalent to include this variable as covariate in the reduced form model. It is also evident that adding a covariate in the random slope model implies to consider a cross level interaction ($X_{ij}C_j$). In this model the only further assumptions needed concern the independence of C_j and the error terms likewise stated for X_{ij} in (3.4) and (3.7).

It is important to see that now the residual is heteroskedastic since its variance depend on the values of the covariate X

$$\begin{aligned} \mathrm{var}\left(\xi_{ij}\right) = & \quad \mathrm{var}\left(e_{ij}\right) + \mathrm{var}\left(u_{0j}\right) + \mathrm{var}\left(u_{1j} X_{ij}\right) + 2\,Cov\left(u_{0j}, u_{1j} X_{ij}\right) \\ & = \theta + \Psi_{00} + X_{ij}^2 \Psi_{00} + 2\,\Psi_{01} X_{ij} \end{aligned} \qquad (3.9)$$

From equation (3.9) we can see that the total residual variance is no longer constant, but is a quadratic function of X. As a consequence the VPC is a function of X_{ij}

$$\tau = \frac{\Psi_{00} + X_{ij}^2 \Psi_{00} + 2\Psi_{01} X_{ij}}{\Psi_{00} + X_{ij}^2 \Psi_{00} + 2\Psi_{01} X_{ij} + \theta} \qquad (3.10)$$

Importantly, in this case the VPC is not equal to the ICC (Goldstein et al, 2002). For random slopes models, it is usual to calculate the VPC as a measure of the importance of clustering for several values of the covariate X_{ij} It could be of specific interest to see how the importance of clustering varies with the X-values.

This model could be generalised with several covariates at the first level, more than two levels and with several covariates at each level, which could be also not necessarily the same. It is usual to insert a random intercept at each level, while it is unusual to consider all the possible random slopes and cross level interactions even if this is theoretically possible. The number of parameters to be estimated raise fast with the number of levels, covariates and random effects included and also the interpretation of parameters could became difficult (Hox, 1995; p. 19). As Di Prete and Forristal (1994) note, we must be aware that our imagination "can easily outrun the capacity of the data, the computer, and current optimization techniques to provide robust estimates".

Researchers that use a multilevel model will include those random slopes and those cross level interactions that seem to be important to insert from a theoretical point of view. We can also test if the inclusion of a random coefficient improves the model fit or not. To test if a model including a random intercept, or a random slope, has to be preferred to a model without it we can employ, when a maximum likelihood estimator is used, a *deviance test*, or *likelihood ratio* test (LR test). We have to note, however, that in this kind of situations we are, basically, testing if the variance of the random effect, to be included or not, is significantly different from zero. Since variances are always positive, the null hypothesis falls on the boundary of the parameter space. As a consequence, the limiting distribution of the LR test statistic is not the usual chi-square with 1 degree of freedom, but is instead a 50:50 mixture of such a distribution and a point mass at 0 (Self and Liang, 1987). To keep this complication into account, Snijders and Bosker (1999) suggest halving the p-value we get using a standard LR test.

From the discussion we made in this section we have seen that multilevel models are very flexible tools that allow exploiting the richness of the

information hidden in a multilevel data structure. We have not posed yet any focus on the estimation of causal effects. In section 3.3.2 we discuss the issue of endogeneity in multilevel linear regression models which is a potential obstacle to causal inference. In the next section, instead, we briefly analyse the role of latent variables in multilevel models and how assign values to the random intercepts and slopes for individual units.

3.3.1 The nature of the latent variables used in multilevel models and how we can obtain their predictions

Latent variables are, in broad terms, unit characteristics which are not directly observed, but are rather inferred from other variables that are surveyed and directly measured. The use of latent variables is common in many fields, such those of social sciences, economics, psychometrics. Skrondal and Rabe-Hesket (2004) note that latent variables are used to represent several phenomena, including true variables measured with error, hypothetical construct, missing data, unobserved heterogeneity.

It is interesting to note that latent variables can be used also to model particular missing data: the counterfactuals. We have already said, in chapter 2, that causal inference can be seen as a missing data problem. In this context, latent variables represent particular missing values, that is, values that would have been realized under counterfactual circumstances, for instance if a treated unit was exposed to control. A particular area where latent variable modelling has been increasingly adopted for causal inference is represented by randomised experiments with noncompliance. For example, Muthén (2002) propose to use latent class models for the estimation of the Complier Average Causal Effect (CACE). We have discussed IV methods for the estimation of CACE, we referred to as Local Average Treatment Effect, in chapter 2.

In the context of multilevel models, latent variables are usually referred to as random effects, and they represent the effect of unobserved covariates not included in the model. It is important to note that unobserved heterogeneity is not a hypothetical construct since it merely represents the combined effect of all

unobserved variables at a given level of the structure. We not give to latent variables in multilevel models any meaning beyond this.

After a multilevel model has been estimated, it is often of interest to assign values to the random effects. The assignment of values to latent variables is referred to as latent scoring (for continuous latent variables) or as classification (for discrete latent variables). Sometimes, scoring and classification are the main aims of latent variable modelling. Examples include disease mapping, small area estimation and assessment of institutional performance. The last approach has been employed to rank organisations (for example schools) as more or less effective according to their random effects scores (see e.g. Aitkin and Longford, 1986). We adopt this approach in chapter 4 in order to compare communities' effectiveness in favouring households to escape poverty. Another important statistical application of latent scoring is in model diagnostics to study the assumptions underlying the model.

There are two main statistical approaches to assign values to the random effects. To formalise this ideas, let us refer to the simple random intercept model in (3.2). In this model we have only one random effect, in particular the random intercept u_{oj}. The first approach is a Maximum Likelihood (ML) procedure, where parameter estimates for the fixed part are assumed as if they were the true parameter values. The ML estimate for u_{oj} are simply obtained as the values that maximize the likelihood of the observed responses Y_{ij}

$$Likelihood\left(Y_{1j},..,Y_{n_j j} \mid X_{1j},..,X_{n_j j}, u_{oj}\right).$$

The second approach is based on *Empirical Bayes* (EB) predictions. In contrast to the ML approach, EB uses information about the prior distribution of the random effects, in addition to the observed responses. In this case the random effects are treated as proper random variables hence the term *prediction* is used in contraposition to ML *estimates*.

The prior distribution of u_{oj} in model (3.2) is just the normal distribution with zero mean and estimated variance plugged in. It represents what we know

about u_{oj} before seeing the data. Once we have observed the responses, the posterior of u_{oj} update our knowledge regarding u_{oj} after seeing the data for cluster j, combining the prior and the likelihood. The EB prediction of u_{oj} is just the mean of the posterior distribution

$$Posterior\left(u_{oj} \mid Y_{1j},..,Y_{n_j j}, X_{1j},..,X_{n_j j}\right)$$
$$\propto P\text{rior}(u_{oj}) \times Likelihood\left(Y_{1j},..,Y_{n_j j} \mid X_{1j},..,X_{n_j j}, u_{oj}\right)$$

where $\propto$ means proportional to. In a linear model with normal error terms, the posterior is normal and then mean is thus equal to the mode. Since the posterior distribution is a compromise between the prior distribution and the likelihood the EB lies between the ML estimates and the mean of the prior. In linear random intercept model the following simple formula relates the EB prediction to ML estimates:

$$\hat{u}_{oj}^{EB} = \hat{R}_j \, \hat{u}_{oj}^{ML}$$

where $\hat{R}_j = \dfrac{\hat{\psi}_{00}}{\hat{\psi}_{00} + \hat{\theta}/n_j}$.

$\hat{R}_j$ is called the *shrinkage factor* because it assume values between 0 and 1, so that the EB prediction is shrunken toward 0, that is, the mean of the prior.

3.3.2 Second level endogeneity in the multilevel linear model

Let suppose that our interest lies in the estimation of β_1 in the model (3.2). For example, the covariate X_{ij} in that model could represent a treatment of interest. In the previous section we stated that in model (3.2) we assume (see assumptions (3.4)) that $E(u_{0j}|X_{ij}) = 0$ and that $E(e_{ij}|X_{ij}) = 0$ which imply, respectively, that $Cov(u_{0j},X_{ij}) = 0$ and that $Cov(e_{ij},X_{ij}) = 0$. In order to consistently estimate β_1 we require that X_{ij} is uncorrelated with both the error terms e_{ij} and $u_{0j.}$ If this

assumption does not hold we speak about endogeneity likewise in the traditional regression model.

In the two levels models we distinguish between two kinds of endogeneity. If a covariate correlates with the error term at the first level we speak about *first level endogeneity*. On the other side if X correlates with the random effects u_0 we speak about *second level endogeneity*. This distinction is not a scholar one but it is important since the two forms of endogeneity can be faced in very different ways. In fact the independence assumption concerning the random effects is not as stringent as it may appear. In fact, as noted for example by Skrondal and Rabe-Hesketh (2004) if the random effects are correlated with a first level variable, such correlation is removed as soon as the cluster mean of such variable is introduced as a further covariate. On the contrary, first level endogeneity, likewise the endogeneity problem in the standard regression model, cannot be solved in a similar easy way but requires the use of specific methods like instrumental variables, simultaneous equation models or others as we discussed in Chapter 2[12].

Here we focus on the problem of the dependency between random effects and a first level covariate. This topic has been investigated in the recent literature (Snijders and Bosker, 1999; Rice et al., 2002;Ebbes et al., 2004; Fielding, 2004; Wooldridge, 2002; Grilli and Rampichini, 2006; Snijders and Berkhof, 2007).

As we mentioned in section 2.3 endogeneity in a regression model often arise due to the omission of relevant variables. Suppose that in our two level linear model (3.2) a covariate that is associated both with the outcome and with the regressor X is excluded then it will be included in the error term which is thus correlated with X implying inconsistent estimation of β_1. The omission of a relevant level-2 covariate will imply, in particular, the second level endogeneity problem. Let explore this issue in more detail. Let suppose that we specify the model

[12] Some authors have implemented specific IV estimators for multilevel models. For example, Spencer and Fielding (2000) extend the Iterative Generalized Least Square estimation procedure to cases where endogeneity is suspected and instrumental variables are available.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + e_{ij} + u^*_{0j} \tag{3.11}$$

whereas the correct model is

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_{0\_1} C_j + e_{ij} + u_{0j} \tag{3.12}$$

In the model (3.11) we omitted the variable *C*, a cluster-level variable implying that the error terms u^*_0 can be written as

$$u^*_{0j} = \beta_{0\_1} C_j + u_{0j} \tag{3.13}$$

If the omitted variable *C* is correlated with *X* then the error term u^*_0 will be correlated with *X* in the (3.11) yielding to inconsistent estimate of the effect of interest, β_1. In econometrics, and in particular in the setting of panel data with unobservable individual heterogeneity, starting from Mundlak (1978), it is usual to model $E(u^*_{0j} \mid X_{ij})$ as a linear function of the cluster mean of X_{ij}, that we indicate with $\overline{X}_j$. Viewed in another way, we can express the dependency between *C* and *X* with the following regression

$$C_j = \alpha_0 + \alpha_1 \overline{X}_j + \omega_j \tag{3.14}$$

We can note that the regression of C_j on $\overline{X}_j$ instead that on X_{ij} is justified by the fact that $X_{ij} = (X_{ij} - \overline{X}_j) + \overline{X}_j$ and the regression coefficient of C_j on $(X_{ij} - \overline{X}_j)$ is 0, since C_j varies only between clusters while $(X_{ij} - \overline{X}_j)$ varies only within clusters. The implication of this reasoning is that including in the model (3.11) the cluster mean $\overline{X}_j$ as a separate covariate will eliminate the dependency between the covariate X and the error term. The model that include both the covariate X_{ij} and its cluster mean $\overline{X}_j$, which can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_{0\_1}\overline{X}_j + e_{ij} + u_{0j} \qquad (3.15)$$

has received specific interest in the literature because $\overline{X}_j$ can represent an important contextual variable and its inclusion allow to separate the so-called between and within clusters effects of the covariate *X*, as was proposed already by Davis et al. (1961). We find useful to deepen this topic.

Distinguishing within and between-group relations is very important in many applied work. In fact, the relationship between a covariate *X* and a dependent variable *Y* within clusters can be completely different from the relationship we can observe between clusters because the processes at work in the two dimensions can be very different. We already mentioned this issue when we discussed in section 3.1 ecological and atomistic fallacy problems. In the model (3.3) we keep into account the multilevel data structure but the parameter β_1 mixes up the between and the within effects of X_{ij} on Y_{ij}. In order to obtain purely between-clusters effects of the explanatory variable we have to average in (3.3) the response and the covariate for each second level unit j over first level units *i* and perform the regression using the resulting cluster means

$$\overline{Y}_j = \beta_0 + \beta^B \overline{X}_j + \overline{e}_j + u_{0j} \qquad (3.16)$$

where β^B indicates the *between regression coefficient*. If, on the other hand, we want purely within-group effects we could subtract the above between-group regression (3.16) from the original model (3.3) to obtain the within model

$$Y_{ij} - \overline{Y}_j = \beta^W \left(X_{ij} - \overline{X}_j\right) + e_{ij} - \overline{e}_j \qquad (3.17)$$

where β^W indicates the *within regression coefficient*, the regression coefficient within each group, here assumed to be the same in each group. The model (3.17) is equivalent to the fixed effect model whose characteristics we discussed in section 3.1, where we stressed the differences between this model and the random effect model. The fixed effect model can be obtained by replacing the

random intercept u_{0j} for each cluster in the original model (3.3) by a fixed intercept α_j. As we already mentioned, this model hamper the inclusion of contextual variables because all group-specific effects are accommodated by α_j, leaving only within-group effects to be explained by covariates. On the other hand, we have to note that this model allows α_j to be arbitrarily correlated with covariates, or said in other words it permits the existence of omitted group-level variables that can be correlated with the observed covariates.

If we sum models (3.16) and (3.17) we get the following multilevel model which allows to study between and within relationships at the same time

$$Y_{ij} = \beta_0 + \beta^B \overline{X}_j + \beta^W \left(X_{ij} - \overline{X}_j \right) + e_{ij} + u_{0j} \tag{3.18}$$

which can be re-written as

$$Y_{ij} = \beta_0 + \left(\beta^B - \beta^W \right) \overline{X}_j + \beta^W X_{ij} + e_{ij} + u_{0j} \tag{3.19}$$

Model (3.19) is just a re-parameterization of model (3.18) but some differences exist between these two models. Model (3.18) could be preferred because the covariates $\overline{X}_j$ and $X_{ij} - \overline{X}_j$ are uncorrelated. Another difference between the two models is conceptual. The use of group-mean centred covariate, likewise in model (3.18), should be motivated, as noted by Snijders and Bosker (1999), "by a clear theory (or an empirical clue) that not in the first place the absolute score X_{ij} but rather the relative score $X_{ij} - \overline{X}_j$ is related to Y_{ij}". These considerations are really important only if we consider random slope models because in this case the model including X_{ij} and $\overline{X}_j$ and the model including $X_{ij} - \overline{X}_j$ and $\overline{X}_j$ are not equivalent (Kreft et al, 1995).

Coming back to the second level endogeneity problem, we note that model (3.19) is exactly the same of model (3.15). It is sufficient to recognize that in model (3.15) the coefficient of X is actually the difference between β^B and β^W, while the coefficient of the cluster mean, $\overline{X}_j$, is equal to the between coefficient,

β^W. Only if the between and within relationships are the same, that is $\beta^B = \beta^W$, then model (3.19) collapse to model (3.3). Hence, we can interpret the second level endogeneity problem as due to model misspecification driven by the erroneous assumption that the between and within regression coefficients are equals. This assumption underlies the basic model (3.3) and, as we have seen, can be avoided simply by including the cluster mean $\overline{X}_j$ as an additional covariate.

In order to test if second level endogeneity exists in a model like (3.3), that is if the second level random effect u_{0j} correlates with X_{ij}, we can use the Hausman specification test (Hausman, 1978), which offers a general procedure to test potential regression model misspecifications. The Hausman test can be used to compare two estimators which are both consistent under the null hypothesis we are testing, in our case the zero correlation between u_{0j} and X_{ij}, while only one of them is still consistent under the alternative hypothesis. Following the general procedure we can use a fixed effects estimator which is consistent under both hypotheses and a random effects estimator which is consistent only under the null. The Hausman test has been criticised (Skrondal and Rabe-Hesketh, 2004) because the rejection of the null hypothesis could be due to model misspecifications different from that we are testing.

A simpler solution in multilevel models consist to test if the within regression coefficient is equal to the between regression coefficient, that could be easily implemented as a Wald test on the coefficient of $\overline{X}_j$ in model (3.19), as suggested by Mundlack (1978). Baltagi (2001) showed that this test is numerically equivalent to the Hausman test.

In this section we considered a simple case of a random intercept multilevel model with a single regressor. However, the resolution of the second level endogeneity problem through the inclusion of cluster-mean covariates applies, *mutatis mutandis*, also in more complicated models, with many endogenous covariates and random slopes (Snijders and Berkhof, 2007).

A complication we have not taken into account up to now, however, is related to the fact that the cluster-mean $\overline{X}_j$ is a *sample* mean used as a

measurement of a *population* mean and hence the model including $\overline{X}_j$ is affected by measurement errors. This issue has been studied by Grilli and Rampichini (2006) who show that the consequences of measurement errors are that neither model (3.18) and (3.19) consistently estimate β^B and the variance at the second level. However, they suggest simple corrections that overcome the problem.

In this section we have not posed any specific attention on the estimation of causal effects but our interest was in presenting the main features of the linear multilevel models. Many authors have found these models very useful for detecting causal effects in observational and randomized. The main attraction of these models is that they allow, as we have seen, to keep into account unobserved cluster effects that impact on the phenomenon under study. In the next section we analyse multilevel linear models under the potential outcome framework in order to see how and if they face the issues introduced in section 3.2.

3.4 Causal inference with traditional multilevel models studied under the potential outcome framework

In the previous section we have seen the principal benefits of using multilevel models. Since our goal in this thesis is to estimate a causal effect, and specifically the causal effect of childbearing events on consumption growth, we find useful to stress the advantages of using multilevel models instead traditional regressions from the prospective of recovering causal effects. Therefore, here we briefly adapt the potential outcome framework meticulously presented in section 2.1 to a multilevel setting.

3.4.1 Adapting the basic notation and definitions

Suppose the population under study is a two-level population consisting of N micro units at the first level, indexed by i ($i = 1, 2, \ldots, n_j$) nested in J macro units at the second level, indexed by j ($j = 1, 2, \ldots, J$). We are interested in the causal effect of a binary variable, D, on a continuous outcome, Y[13]. Under SUTVA, each unit, ij, has two potential outcomes depending only on its assignment to the treatment levels: Y_{1ij} if $D_{ij} = 1$ and Y_{0ij} if $D_{ij} = 0$. As we have seen in section 2.1, SUTVA implies that the potential outcomes for any unit do not vary with the treatments assigned to any other unit. In particular, in a multilevel setting, SUTVA implies that potential outcomes for unit ij are not influenced by treatments received by units belonging to other clusters and even belonging to the same cluster j. We discuss the potential violation of SUTVA and some weaker assumptions we can formulate in a multilevel setting in the following section 3.5. Traditional multilevel models implicitly assume SUTVA. Hence, in the remaining of this section we maintain this assumption.

Similarly to what we did in section 2.1.2, we define the causal parameters ATE, ATT and ATU:

$$\text{ATE} = E(Y_{ij1} - Y_{ij0}) \tag{3.20a}$$

$$\text{ATT} = E(Y_{ij1} - Y_{ij0} \mid D_{ij} = 1) \tag{3.20b}$$

$$\text{ATU} = E(Y_{ij1} - Y_{ij0} \mid D_{ij} = 0) \tag{3.20c}$$

where the expectations are taken, respectively, on the whole population, on the population of treated and on the population of untreated without keeping into account that units are nested in clusters.

In multilevel models it is usual to employ variables measured at each level of the hierarchy as covariates. We indicate with X a covariate measured at the first level and with C those measured at the second level (contextual variables). The first level covariates are included in the set $\boldsymbol{X}$, while the contextual covariates are included in the set $\boldsymbol{C}$. Multilevel models can also

[13] The hypotheses that the hierarchy is two-level, treatment is binary and outcome is continuous are maintained for simplicity in all the discussion but can be removed to allow for a more general context.

include, as aforementioned, cluster mean averages of variable measured at the first level, both to control for second level endogeneity and to study potentially interesting compositional effects.

3.4.2 Studying some multilevel models

We can derive a simple two-level random intercept model in the following way. Likewise we made in section 2.3.1 let's specify two multilevel models for each potential outcome:

$$\begin{aligned} Y_{0ij} &= X_{ij}\beta + C_j\theta + e_{ij} + u_{0j} \\ Y_{1ij} &= X_{ij}\beta + C_j\theta + \Delta + e_{ij} + u_{0j} \end{aligned} \tag{3.21}$$

Then the model for the observed outcome is

$$\begin{aligned} Y_{ij}^{obs} &= \left(X_{ij}\beta + C_j\theta + \Delta + e_{ij} + u_{0j}\right) * D_{ij} + \left(X_{ij}\beta + C_j\theta + e_{ij} + u_{0j}\right) * \left(1 - D_{ij}\right) \\ &= X_{ij}\beta + C_j\theta + D_{ij}\Delta + e_{ij} + u_{0j} \end{aligned} \tag{3.22}$$

In the previous model is assumed that the effect of the treatment is constant, that is it does not vary between units. In this case ATE = ATT. Importantly, the fundamental identifying hypothesis underlying the previous model is a modified version of unconfoundedness:

$$Y_1,\ Y_0 \perp D \mid \boldsymbol{X},\ \boldsymbol{C},\ u_0 \tag{3.23}$$

The previous assumption, that we can call "multilevel unconfoundedness" assumes that if we simply condition on observed covariates $\boldsymbol{X}$ and $\boldsymbol{C}$ some dependency among potential outcome still remains and it is caused by unobserved factors at the cluster levels, which we represent by a latent variable u_0[14]. Hence, controlling also for u_0 this dependency is cancelled out.

[14] We have to note that the latent variables we use in multilevel models have a particular nature, as happens for all latent variables. They are not simply unobserved variables. They are assumed to be zero mean variables, uncorrelated with covariates and, in a sense, include the effect of the unobserved variables. In fact, we have not a coefficient for those variables in the model.

This is a weaker assumption with respect to the unconfoundedness assumption stated in section 2.3 (see assumption 2.31).

As we said in the section 2.3.1 we can simply overcome the constant treatment effect assumption allowing the treatment effect to change with respect to all covariates obtaining a Fully Interacted Linear Model. This is obtained allowing the coefficient of covariates to be different in the two models for potential outcomes:

$$\begin{aligned} Y_{0ij} &= X_{ij}\beta + C_j\theta + e_{ij} + u_{0j} \\ Y_{1ij} &= X_{ij}\beta^1 + C_j\theta^1 + \Delta + e_{ij} + u_{0j} \end{aligned} \tag{3.24}$$

$$\begin{aligned} Y_{ij}^{obs} &= \left(X_{ij}\beta^1 + C_j\theta^1 + \Delta + e_{ij} + u_{0j}\right) * D_{ij} + \left(X_{ij}\beta + C_j\theta + e_{ij} + u_{0j}\right) * \left(1 - D_{ij}\right) \\ &= X_{ij}\beta + C_j\theta + D_{ij}\Delta + I_{Xij}\phi + I_{Cij}\varphi + e_{ij} + u_{0j} \end{aligned} \tag{3.25}$$

where the matrixes $\boldsymbol{I_X}$ and $\boldsymbol{I_C}$ includes, respectively, all the interactions among $\boldsymbol{X}$ and D and among $\boldsymbol{C}$ and D. The vectors Φ and φ collect, respectively, the coefficients of variables included in $\boldsymbol{I_X}$ and $\boldsymbol{I_C}$. They coincide with the difference between the correspondent vectors of coefficients in the model for the two potential outcomes (3.24): $\Phi = \beta^1 - \beta$ and $\varphi = \theta^1 - \theta$. In this case ATE and ATT are, in general, different. We also note that in this case the parameter Δ does not represent the ATE as in model (3.22) but it represents the effect of D when the variables interacting with it are all equal to zero. The ATE and ATT in this model are, respectively, given by

$$\begin{aligned} ATE &= \left(\beta^1 - \beta\right)'E(X) + \left(\theta^1 - \theta\right)'E(C) + \Delta \\ ATT &= \left(\beta^1 - \beta\right)'E_{T=1}(X) + \left(\theta^1 - \theta\right)'E_{T=1}(C) + \Delta \end{aligned}$$

However, the inclusion of all the interactions in a multilevel model is uncommon. On the contrary, the usual way used in multilevel modeling to allow the effect of a variable to vary is to include a random slope. In our context this

implies that the effect of the treatment differs by second-level clusters. We can derive a random slope model in the following way

$$\begin{aligned} Y_{0ij} &= X_{ij}\beta + C_j\theta + e_{ij} + u_{0j} \\ Y_{1ij} &= X_{ij}\beta + C_j\theta + \Delta + e_{ij} + u_{0j} + u_{1j} \end{aligned} \tag{3.26}$$

$$\begin{aligned} Y_{ij}^{obs} &= \left(X_{ij}\beta + C_j\theta + \Delta + e_{ij} + u_{0j} + u_{1j}\right) * D_{ij} + \left(X_{ij}\beta + C_j\theta + e_{ij} + u_{0j}\right) * \left(1 - D_{ij}\right) \\ &= X_{ij}\beta + C_j\theta + D_{ij}\Delta + e_{ij} + u_{0j} + u_{1j}D_{ij} \end{aligned} \tag{3.27}$$

Hence, we can obtain a random slope model by allowing the error terms in the two models to have a different structure. Each unit *ij* is subject to cluster effects depending on the cluster *j* to which it belongs and represented by latent variable shared by each unit in the cluster. However, two latent variables at the cluster level are in action: one, u_0, influences both potential outcomes and hence influences the baseline level of Y^{obs}. This latent variable represents all combined effect of unobserved community level variables that influence the level of the outcome but do not impact on the effect of the treatment. The other latent variable, u_1, instead, influences only potential outcomes of the treated units, representing unobserved factors that modify the *effect* of the treatment. In fact, the effect of the treatment for units belonging to the cluster *j* is $Y_1 - Y_0 = \Delta + u_1$. As a consequence, we have an interaction between u_1 and D in the model for the observed outcome, Y^{obs} (see the 3.27). In this setting the ATE and ATT are in general different because we could have more treated in communities where treatment has a higher effect or vice versa. In this model ATE and ATT are given by

$$\begin{aligned} ATE &= \Delta + E(u_1) = \Delta \\ ATT &= \Delta + E_{T=1}(u_1) \end{aligned}$$

Since, by construction, u_1 is a random variable with zero mean, ATE coincides with the parameter Δ in the model (3.27). ATT is different from ATE if $E_{T=1}(u_1)$ is different from 0.

Conditioning to one specific cluster ATE and ATT coincide. It could be of interest to calculate conditional ATEs:

$$ATE(j) = E(Y_1 - Y_0|j) \tag{3.28}$$

The parameter in (3.28) is the causal effect of the treatment for units belonging to the specific cluster *j*. It can be estimated using a "fixed effects" approach implementing separate regressions in each cluster or including dummies for cluster. Following a random effects approach, proper of multilevel models, we can estimate (3.28) by first estimating parameter Δ and then adding the a-posteriori prediction of the error term u_{1j}, using multilevel model (3.27). In this way parameter (3.28) coincides with the following parameter:

$$ATE(u_1) = E(Y_1 - Y_0|u_1) = \Delta + u_1 \tag{3.29}$$

In (3.29) it is made clear that, as in the traditional way to approach multilevel models, we are basically interested in the variance of the random variable u_1 and not in the cluster-specific estimates. If this variance is significant, then the causal effect of the treatment changes considerably by cluster. In this way we can generalise results to the population of clusters, of which observed clusters represent a sub-sample: clusters with similar values of u_1 have similar values also for the causal effect.

In the previous models we assumed that *D* is exogenous. If we assume that only first-level exogeneity holds but *D* is correlated with the error term at the second level, giving rise to the second level endogeneity, we have to modify the unconfoundedness assumption. In other words, we could suspect that some relevant covariate at community level is unobserved and hence UNC would hold only conditional also on this covariate *C**:

$$Y_1,\ Y_0 \perp D \mid \boldsymbol{X},\ \boldsymbol{C},\ C^*,\ u_0 \tag{3.29}$$

Obviously, we cannot condition on C^* but we can think to substitute it with $\overline{D}_j$, the cluster mean of D_{ij} as done in the traditional multilevel model (and explained before in section 3.2). The UNC could be expressed as

$$Y_1,\ Y_0 \perp D \mid \boldsymbol{X},\ \boldsymbol{C},\ \overline{D}\ ,\ u_0 \tag{3.30}$$

and the model becomes:

$$\begin{aligned} Y_{0ij} &= X_{ij}\beta + C_j\theta + \overline{D}_j\alpha + e_{ij} + u_{0j} \\ Y_{1ij} &= X_{ij}\beta + C_j\theta + \Delta + \overline{D}_j\alpha + e_{ij} + u_{0j} \end{aligned} \tag{3.31}$$

$$\begin{aligned} Y_{ij}^{obs} &= \left(X_{ij}\beta + C_j\theta + \Delta + \overline{D}_j\alpha + e_{ij} + u_{0j}\right)* D_{ij} \\ &+ \left(X_{ij}\beta + C_j\theta + \overline{D}_j\alpha + e_{ij} + u_{0j}\right)* \left(1 - D_{ij}\right) \\ &= X_{ij}\beta + C_j\theta + D_{ij}\Delta + \overline{D}_j\alpha + e_{ij} + u_{0j} \end{aligned} \tag{3.32}$$

where $\overline{D}_j$ is the cluster mean of Δ and represents the proportion of treated in the cluster. The (3.32) is the traditional multilevel model where we control for second level endogeneity by including the cluster means of the endogenous regressor. In this case, as in model (3.22), ATE is equal to ATT and coincides with the parameter Δ.

However, the parameter Δ in model (3.32) is, in general, expected to be different from Δ in model (3.22). This is because now the coefficient of D estimates the within-cluster effect of the treatment while in model (3.22) it mixes up the within and the between effects. Model (3.31) highlights an interesting fact: in order to disentangle the within and the between effects in the model for observed outcome (3.32), or, equivalently, in order to control for second level endogeneity of D, we had to include the cluster mean $\overline{D}_j$ in the models for potential outcomes. This clearly violates the SUTVA because we are writing the potential outcomes for unit *ij* as depending on the proportion of treated in the

cluster *j*, to which it belongs. Interestingly, we note that in multilevel settings the traditional consideration about the need of disentangling within and between effects is closely related to the potential violation of the standard SUTVA assumption.

Summarizing, in this section we analyzed the traditional multilevel models under the standard potential outcome framework clarifying that we can use these models to recover causal effects under specific assumptions. Basically, the identifying assumptions used in this section correspond to modified versions of unconfoundedness which are weaker with respect to the version presented in section 2.3 (assumption 2.31). However, we noted that sometimes the parameters of multilevel models are not directly corresponding to the standard parameters of interest in causal inference (ATE and ATT).

Moreover, if we suspect the within and between effects to be different, and this is the most likely situation in general, we have to specify a more sophisticated model which keeps into account the violation of SUTVA and hence, needs to recognize that more potential outcomes are in action (see the third point in section 3.2). This is not only a methodological complication but gives also us the opportunity to discover further potentially interesting causal parameter. Another issue, we have not posed in this section, is related to the second point presented in section 3.2. We face these topics in the next section.

3.5 Causal inference under the potential outcomes framework in a multilevel setting

In this section we discuss the issues outlined in section 3.2. First, we briefly discuss the literature about the topic. The literature about causal inference in observational multilevel settings, traditionally, has focused on the use of the multilevel models to estimate causal effects in the considerations of the benefits of this kind of models. However, in this part of the literature no care has been put on complications that arise in multilevel models, such as the violation of SUTVA. Only using an appropriate conceptual framework to study causal

analysis, such as the potential outcomes one, we can appropriately address such problems of causal inference.

The literature on causal inference under the potential outcome framework extended to a multilevel setting is quite limited. Moreover, some of these works relate on the estimation of cluster effects (Sobel, 2006; Oakes, 2004; Subramanian, 2004; Diez-Roux, 2004; Gitelman, 2005; Harding, 2003; Stuart, 2007). Only few works address the problem of estimate the causal effect of treatment received by units clustered in macro units: Kim and Seltzer (2007), Hong (2003), Hong and Raudenbush (2005 and 2006). This literature is lacking to address simultaneously the three issues outlined in section 3.2. Kim and Seltzer (2007) explicitly consider the cluster-specificity of the selection process (the second point in section 3.2), while Hong and Raudenbush (2005 and 2006) address the problem of the violation of the SUTVA.

For clarity we re-write the three issues outlined in section 3.2 here:

1. Cluster-heterogeneity of the treatment effect,
2. The multilevel nature of the selection process,
3. Potential violation of the SUTVA.

3.5.1 Cluster-heterogeneity of the treatment effect

Actually, the first issue is not a statistical one but it is driven from a research question about the heterogeneity of the treatment effect. Moreover, it is not specific to multilevel settings. In all studies of causal inference we could be interested in the treatment effect heterogeneity. For example, in our case we would like to know if the effect of childbearing events is different for farmers and non-farmers; for kinh versus non-kinh households; for households with different educational levels and so on. In a multilevel setting, to these aspects we sum another that is specific to this kind of context, consisting on the fact that the heterogeneity can be driven by the characteristics of the cluster to which units belong. In our case, it could be due to the place where households live.

As we mentioned in section 3.2, the issue we want to analyse at the first point refers to the fact that the effect of the treatment can be higher in some clusters and lower in others. Then, we could be interested in the estimation of cluster specific causal effects like parameter (3.28). If we suspect that the cluster-heterogeneity is due to the interaction between the treatment and a specific and observed cluster variable we could estimate causal effects conditional on different values of this variable, C: $E(Y_1 - Y_0 | C=c)$. For example, in our case C could represent the presence of a specific infrastructure that in some way helps households with children (health care centres, family planning centres, etc.). In this case we would estimate the effect of childbearing events in communities with and without the infrastructure. This is important for policy making. In fact, if the effect of childbearing events is negative in all kind of communities but has a lower magnitude in communities with the infrastructure, this is a clear indication that expanding the presence of this facility can ameliorate the conditions of households having children. If we do not have a clear theoretical indication of which community variable C is expected to interact to the treatment indicator, or if this variable is unobserved, then we could be interested in the estimation of parameter (3.29). Actually, we are interested in the variance of u_1. If it is significative, we can argue that some unobserved community characteristics make the treatment to have a stronger effect in some communities and a weaker one in others. A qualitative analysis, then, could help to understand which characteristics drive this heterogeneity.

From a methodological point of view, to address the study of the cluster-heterogeneity in the treatment effect we could employ a multilevel model with a random slope for the treatment indicator such as model (3.27). Likewise single level regression models, the model (3.27) impose a linear relationship between the outcome and the treatment. Moreover, we cannot control the balancing process as we do when we use matching methods (see chapter 2). As a result, we think that a better strategy is to first implement a matching method. Then, if the estimated effect is significant, we could run a multilevel model like (3.27) on the matched sample to study cluster-heterogeneity in the treatment effect. Actually, if we run this model on the matched sample we can avoid the inclusion of

additional control covariates. However, these can be replaced by the estimated propensity score in order to further control for observed covariates that could be not perfectly balanced in the treated and control groups.

3.5.2 The multilevel nature of the selection process

When we implement the propensity score matching in a multilevel setting, we should recognise that, not only the outcome model, but also the selection process can have a multilevel structure. This is the case when the probability of being treated changes substantially by cluster and the effect of some covariate on this probability varies by cluster. The first aspect requires the inclusion of a random intercept in the model of the propensity score, while the second one asks for the inclusion of random slopes. In other words, the statistical implication of the multilevel structure of the selection process is that including only observed covariates in the model for the propensity score, and hence balancing only for them, could not be sufficient. Some unobserved cluster level characteristic could be related to both the treatment and the outcome, generating bias in the estimation of causal effects. Therefore, we propose to use a methodology ensuring a balancing between treated and controls also for the unobserved clusters factors.

Kim and Seltzer (2007) proposes to use a multilevel model for the estimation of the propensity score and then to implement the matching algorithm within each cluster. If we impose that treated and matched controls must belong to the same cluster, then we achieve automatically a perfect balancing in all the observed and unobserved cluster characteristics. However, using a multilevel model for the propensity score have some advantages as described in Kim and Seltzer (2007). This strategy is not feasible in those situations, constituting the norm in social observational studies, where we have few units within each cluster. In these cases, in fact, is likely that in several clusters is difficult to found for each treated a good matched control. This is also our situation since in our application the number of households in a community ranges from 7 to 21.

We propose an alternative procedure consisting of these two stages:

1. Estimate a multilevel model for the propensity score and obtain prediction of the random effects,
2. Estimate a single level model for the propensity score including as additional covariates the predictions of the random effects obtained at the previous stage.

In the first stage we estimate a propensity score model which includes a random intercept, as well as some random slopes. This model has the following form:

$$g(\pi_{ij}) = X_{ij}\beta + C_j\theta + u_{0j} + X_{ij}U \tag{3.33}$$

where g represents the link function, usually the probit or logit function, π is the probability of receiving the treatment and U is a K x 1 column vector containing the random variables u_1, ..., u_k, representing the random slopes. Obviously, the employed multilevel model will contain only a limited number of random slopes. In this case, some elements of U are set to zero. After estimating this model we can obtain empirical bayes predictions of the random effects, as the mean of their respective posterior distributions. These are included as additional covariates in the estimation of a single level propensity score model:

$$g(\pi_{ij}) = X_{ij}\beta + C_j\theta + \hat{u}_{0j}\gamma + \hat{U}_j\vartheta \tag{3.34}$$

Probabilities estimated through model (3.34) will be employed in the matching procedure. Matching on a model like (3.34) can ensures the balancing of all observed covariates at the first and second level, as well as the balancing of random effects. Since random effects capture unobserved cluster characteristics this strategy allows us to balance also these. This aspect can be very important especially in those situations where no information is available at the cluster level. It is interesting to note that also the inclusion of random slopes in model (3.33), and hence their empirical bayes predictions in model (3.34), can be important, in particular, for those confounders whose effects vary substantially

by cluster. By matching on a model like (3.34) we not only balance the covariates but also the effects of covariates with random slopes. We investigate these issues in the application will be presented in chapter 6.

We note that the two-stage procedure here outlined could be inefficient with respect to obtaining directly from model (3.33) the predictions of probabilities plugging in the predictions of the random intercept:

$$\hat{\pi}_{ij} = g^{-1}\left(X_{ij}\hat{\beta} + C_j\hat{\theta} + \hat{u}_{0j} + X_{ij}\hat{U}\right) \tag{3.35}$$

However, since the focus here is on the balancing we can achieve with the different strategies, this loss of efficiency is not matter of concern. Anyway, predicted probabilities in (3.35) should be similar to those obtained from the two-stage procedure. However, we argue that the two-stage procedure it is appealing since it can be seen as a procedure similar to a sensitivity analysis we carry on in chapter 5. This is a sensitivity analysis to violation of the UNC which, basically, consists to simulate an unobserved confounder and put it into the propensity score procedure. Then, the estimates obtained matching only on the original covariates and on covariates plus the simulated confounder are compared, to assess their sensitivity. Our two-stage method mimics this procedure, in the fact that here we compare the estimate we get matching on probabilities estimated through (3.34) and the estimate we obtain by using probabilities estimated with a single level model. If the estimates are similar we can conclude that the PSM is not dramatically sensitive to departure from the UNC due to unobserved cluster effects. Moreover, we suspect that the two-stage procedure is more robust to miss-specification of model (3.33). For example, a second level endogeneity problem in model (3.33), concerning one or more first level covariates, can be faced, as explained in section 3.3.2 by including the cluster mean of the endogenous covariates. However, the two-stage procedure could relax this problem. Also this issue is material for future research.

It is important to note that the probabilities estimated through the (3.35) are, in general, different from the predicted probabilities which are commonly

obtained in multilevel analysis. These are the so-called empirical bayes predicted probabilities and can be obtained as:

$$\hat{\pi}_{ij}^{EB} = \int g^{-1}\left(X_{ij}\hat{\beta} + C_j\hat{\theta} + \hat{u}_{0j} + X_{ij}\hat{U}\right) \times Posterior\left(u_{oj} \mid Y_{1j},..., Y_{n_j j}\right) du_{oj} \quad (3.36)$$

We reserve to future work a more formal comparison among these three different ways to obtain the propensity score estimates (through the (3.35), the (3.36) or the two-stage procedure). In particular, we are planning to implement a simulation study to verify under which condition the three procedures are expected to give sensibly different estimated probabilities and dissimilar balance in covariates and random effects in the treated and control groups.

3.5.3 A weaker version of the SUTVA

The third problem relates to the potential invalidity of the SUTVA in a multilevel setting. The reasons that make us suspect this assumption is untenable in a multilevel setting can be several and depend on the specific studied context and phenomenon. In general this assumption is problematic when sharing and competition for resources generates interference among units (at least) belonging to the same cluster. We will discuss some source of violation of SUTVA in our context when presenting the application in chapter 6.

If we do not use the SUTVA as done up to now, each unit *ij* has not simply two potential outcomes because these depend also on the treatments received by the other units. In general, without SUTVA potential outcomes for each units *ij* depend on the entire *N* x 1 vector of treatments indicator, $\boldsymbol{D}$. Therefore, each unit has 2^N potential outcomes depending on which treatment it receives and on which treatments receive the remaining *N-1* units in the population: $Y_{ij}(\boldsymbol{D}) = Y_{ij}(D_{ij}, \boldsymbol{D}_{-ij})$; where $\boldsymbol{D}_{-ij}$ represent the treatments received by all units in the population except *ij*. Any contrast between two of the 2^N potential outcomes define a causal parameter.

In a multilevel framework it is usual to assume that SUTVA holds at the cluster level even if it is violated within cluster. In fact, a way to overcome the

potential violation of SUTVA is to choice the minimum aggregate level for which we can reasonably state this assumption. This is the traditional way to handle the problem (see e.g. Stuart, 2007). However, the consequence is that the analysis should be conducted at an aggregate level and we cannot refer our results to the individual level. Otherwise we could make an ecological fallacy error, as discussed in section 3.1. Since in our application, as it is often the case in multilevel analyses, we are interested in drawing inference at the unit level we need a weaker version of SUTVA that allow us to continue to run the study at the first level.

If we assume no interference among clusters, that is that SUTVA holds at the cluster level, we already obtain a sensible reduction in the potential outcomes. In fact, in this way the potential outcomes for each unit *ij* depend only on the units belonging to the same cluster *j:* $Y_{ij}(\boldsymbol{D}) = Y_{ij}(D_{ij}, \boldsymbol{D}_{-ij}^{(j)})$; where $\boldsymbol{D}_{-ij}^{(j)}$ represents the treatments received by all units in the cluster *j* except the unit *ij*. Consequently, the potential outcomes for each individual *ij* belonging to the cluster *j* are 2^{n_j}; where n_j is the number of units belonging to the cluster *j*. Anyway, also in this case the potential outcomes are too much and implementing a study of causal inference is difficult. For example, with clusters all of size equal to 10 the potential outcomes are $2^{10} = 1024$ and they fast increase with the cluster size. Moreover, if the clusters, as usual, have different size the number of potential outcomes differs by cluster. This situation makes difficult the definition and interpretation of causal effects requiring to conveniently summarize the vector $\boldsymbol{D}_{-ij}^{(j)}$.

In most situations potential outcomes for a given unit can be thought as influenced by how much units in the clusters receive the treatment while is not important who these units are. As a consequence, the relevant information contained in the vector $\boldsymbol{D}_{-ij}^{(j)}$ is summarized by the proportion of treated units in the cluster (calculated excluding the unit *ij*) that we indicate with $P_{-ij}^{(j)}$. Therefore, potential outcomes for unit *ij* can be written as a function of the

treatment the unit receive and the proportion of the other units treated in the cluster: $Y_{ij}(\boldsymbol{D}) = Y_{ij}(D_{ij}, P^{(j)}_{-ij})$.

For further simplifying the discussion and make inference treatable, we can split the range of $P^{(j)}_{-ij}$ in a limited number of intervals and assume that interference among units belonging to the same cluster is fully captured by these intervals. The information contained in $P^{(j)}_{-ij}$ will be summarised by a scalar function, f, taking s values:

$$f\left(P^{(j)}_{-ij}\right) = \begin{cases} 0 & \text{if } 0 \le P^{(j)}_{-ij} < t_1 \\ 1 & \text{if } t_1 \le P^{(j)}_{-ij} < t_2 \\ \dots & \dots \\ s-1 & \text{if } t_{S\text{-}1} \le P^{(j)}_{-ij} < 1 \end{cases}$$

where s is a positive integer; $t_1, t_2, \dots, t_{s-1}$ are real numbers satisfying: $0 < t_1 < t_2 < \dots < t_{s-1} < 1$. In this way, potential outcomes can be written know as $Y_{ij}(\boldsymbol{T}) = Y_{ij}(T_{ij}, f(P^{(j)}_{-ij}))$.

It is convenient for practical reasons to substitute $P^{(j)}_{-ij}$ with the proportion of treated in the cluster (calculated including unit ij), indicated by P_j. This is not problematic if we can assume that the treatment received by a single unit cannot significantly modify the proportion of treated in the cluster.

The simplest situation, with the minimum number of potential outcomes, is obtained when we fix k at two. In this case, we divide the clusters in those with a "high" proportion of treated and those with a "low" proportion of treated. Let represent with L_j the binary indicator taking value 1 if the proportion of treated in cluster j is "high" and 0 otherwise. In this case, the potential outcomes for unit ij are: $Y_{ij}(\boldsymbol{D}) = Y_{ij}(D_{ij},L_j)$. Hence, we have only 4 potential outcomes according to the treatment the unit receive and to the level of proportion of treated in the cluster:

Y_{11} if $D_{ij} = 1$ and $L_j = 1$

Y_{10} if $D_{ij} = 1$ and $L_j = 0$

Y_{01} if $D_{ij} = 0$ and $L_j = 1$

Y_{00} if $D_{ij} = 0$ and $L_j = 0$.

These 4 potential outcomes are defined under a weaker version of the SUTVA, with respect to the standard one, that we can summarize as follows:

(A weaker version of SUTVA)

$$Y_{ij}(\mathbf{D}) = Y_{ij}(D_{ij},L_j); \tag{3.37}$$

in words, this amounts to assume that there is no interference among units belonging to different clusters, while the within-cluster interference is fully captured by the level of the proportion of treated (high versus low).

Each contrast between two of the 4 potential outcomes define a causal parameter of potential interest. We can conveniently think to this context as we had two treatments: one, *D*, working at the first level and the other, *L*, working at the second level.

A first group of causal parameters of potential interest is given by:

$$ATE^{D}_{|L=1} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid L=1) \tag{3.38}$$
$$ATE^{D}_{|L=0} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid L=0) \tag{3.39}$$

$$ATT^{D}_{|L=1} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid D=1, L=1) \tag{3.40}$$
$$ATT^{D}_{|L=0} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid D=1, L=0) \tag{3.41}$$

The parameters (3.38) and (3.39) measure, respectively, the average causal effect of the treatment *D* in clusters with high proportion of treated ($L = 1$) and with low proportion of treated ($L = 0$). The parameters (3.40) and (3.41) are the correspondent versions of parameters (3.38) and (3.39) calculated conditioning on the sub-group of units with $D = 1$. We can obtain the marginal version of these two parameters as a weighted average of the conditional parameters:

$$ATE^{D} = ATE^{D}_{|L=1} * P(L=1) + ATE^{D}_{|L=0} * P(L=0) \ , \quad (3.42)$$

$$ATT^{D} = ATT^{D}_{|L=1} * P(L=1 \mid D=1) + ATT^{D}_{|L=0} * P(L=0 \mid D=1) \quad (3.43)$$

From the analysis of these parameters we see that the problem of the violation of SUTVA conducts us to consider some interesting new causal estimands. Under the weaker version of SUTVA that we have introduced, we are naturally asked to answer if the effect of treatment is different in cluster where the proportion of treated is low with respect to clusters where this proportion is high. This can be a very interesting comparison, useful for policy making. Moreover, we have to note that parameters (3.41) and (3.42) obtained under this weaker version of SUTVA are not, in general, equivalent to the corresponding parameters calculated under its standard version, as parameters defined in chapter 2. In fact, in general, ATE and ATT defined under SUTVA will confuse the effect of *D* with the effect of *L*. This consideration is similar to the reasoning we made in section 3.3 about the need for disentangle within and between effects in multilevel models. Interestingly, in section 3.4 we have already noted that violation of SUTVA and the within/between effects difference are closely related issues.

Parameters (3.38)-(3.43) estimate the effects of the treatment *D*. In an analogous way we can define similar parameters estimating the effect of the treatment *L*. For example, the corresponding versions of the parameters (3.42) and (3.43) are:

$$ATE^{L} = ATE^{L}_{|D=1} * P(D=1) + ATE^{L}_{|D=0} * P(D=0) \quad (3.44)$$

$$ATT^{L} = ATT^{L}_{|D=1} * P(D=1 \mid L=1) + ATT^{L}_{|D=0} * P(D=0 \mid L=1) \quad (3.45)$$

At this point we have to clarify under which assumption we can identify the parameters we have here introduced and which estimating method we can use. As we have already said, we can treat this situation as the case in which we have two different treatments. Imbens (2000), Lechner (2001) and Cuong (2007) analyze the case of causal inference in the presence of multiple treatments under

the potential outcome framework. Building on Cuong (2007), we can state the identifying assumptions for our case as follows:

$$(Y_{11}, Y_{10}, Y_{01}, Y_{00}) \perp (D, L) | X, C; \quad (3.45)$$

$$0 < \mathrm{P}(D=1|X, C, L) < 1 \; ; \; 0 < \mathrm{P}(L=1|X, C, D) < 1. \quad (3.46)$$

Assumptions (3.45) and (3.46) are, basically, a generalization of assumptions (2.31) and (2.32) we have seen in chapter 2 for the case of one single treatment. If we are not interested in all the parameters (3.38)-(3.45) we can use weaker versions of assumptions (3.45) and (3.46). For example, if our interested lies in $ATE^{D}_{|L=1}$, then we need only that $(Y_{11}, Y_{01}) \perp D \mid X, C$ and $0 < \mathrm{P}(D=1|X, C, L=1) < 1$.

In order to estimate causal parameters (3.38)-(3.45) we can use a PSM procedure. For the estimation of parameters (3.38)-(3.43) we need first to estimate two propensity score models: $P(D=1|X, C, L=1)$ and $P(D=1|X, C, L=0)$. The former will be employed in the matching algorithm for the estimation of parameter (3.38) or (3.40). The latter, will be employed for the estimation of parameters (3.39) or (3.41). Parameters (3.41) and (3.43) will be estimated through their conditional versions. The previous discussion about the need of considering the multilevel nature of the selection process is still valid. Therefore, the two propensity score models can be estimated using the two-stage strategy discussed before.

Chapter 5
Estimation results of the causal effect of fertility on poverty in Vietnam

Introduction

The estimation of the effects of demographic events on households' living standards introduces a range of statistical issues. In this chapter we analyze this topic, considering our observational study as a quasi-experiment, in which the treatment is expressed by childbearing events between two time points (the two waves of the VLSMS) and the outcome is the change in the equivalized household consumption expenditures.

The chapter is organised as follows. Section 5.1 briefly recalls the basic motivations underlying the current application, and presents some simple descriptive statistics showing a clear negative association between consumption expenditure and the number of children. In Section 5.2 we assume that the unknown assignment mechanism in our observational study is regular. Therefore we present and discuss results from the estimation of methods relying on the UNC assumption. In section 5.3, we assess the robustness of the PSM. In the section 5.4 we propose two instruments for the identification of the causal effect of interest, avoiding the UNC assumption. In this case a latent regular assignment mechanism is assumed. Results from the estimation of IV methods are presented and discussed in section 5.5. Section 5.6 concludes.

5.1 Motivations

In chapter 1, we extensively, described the background and motivations of the application we show in the present chapter. Our main question concerns the estimation of the causal effects of fertility on households' economic wellbeing. Fertility is measured by childbearing events, which are generally believed to be endogenous with respect to consumption expenditure, which is the most common way to measure a households' economic wellbeing. Economic wellbeing is measured in terms of the change in consumption expenditure between the two waves. As discussed in chapter 1, the relationship between fertility and poverty is country-specific. In fact, previous work found mixed results, but the common observation in many LDC is that households with more children show worse living standards. Table 5.1 seems to confirm this negative relationship between fertility and economic wellbeing also for Vietnam.

From table 5.1 we can see that the higher was the number of new children born in the household between the two waves he lower was the consumption expenditure growth in the same period. Of course, simple descriptive statistics like those presented in Table 5.1 do not say anything about causality. Rather, they merely show a negative *association* between number of children and consumption expenditure growth. In order to draw causal conclusions, we have to consider that households with more children are different from households with fewer children with respect to a range of factors. In other words, there is self-selection in the level of childbearing events, which depend on characteristics that are likely to be associated also to consumption, and hence confound the fertility-poverty relationship. Therefore, we cannot simply compare the consumption expenditure growth for households with different level of childbearing events. We need to use adequate methods for causal inference, such those discussed in chapter 2.

Table 5.1: Average equivalized household consumption expenditures at the two waves and its growth by number of children born between the two waves.

N. of children born between the two waves	Observations	Average consumption in 1992	Average consumption in 1997	Average consumption growth in 1992-1997
0	968	970	2436	1466
1	407	856	1892	1036
2	138	790	1755	965
3	24	571	1154	583
At least 1	569	832	1835	1004
Total	1537	916	2201	1285

Notes: We consider the number of children of all household members born between the two waves. All consumption measures are valued in dongs and rescaled using prices in 1992. The 1537 households represented in the table are selected taking only households with at least one married woman aged between 15 and 40 in the first wave. Consumption is expressed in thousands of dongs.

Let's formalise the structure of our quasi-experiment. We have a sample of households under study indexed by $i = 1, 2, \ldots N$, a treatment indicator D that assumes the value 1 for treated units (in our case, for households in which was born at least one child between the two waves) and 0 for untreated or the controls (that is, households not having new children) and an outcome variable, which is the growth in household equivalized consumption between the two waves, indicated by Y.

Our sample is restricted to households where in the first wave there was at least one married woman aged between 15 and 40 years. This selection could be thought as a part of the whole matching strategy. In this way, in fact, we avoid to compare households who would not be eligible for the treatment (that is, not a risk to have children), because they had no woman in fecund age. Obviously different selection strategies are possible[15].

[15] We tried the following alternative selection criteria: 1) households with at least one married woman aged 15-35 in the first wave; 2) households whose head or its spouse is a married woman aged 15-40 in the first wave; 3) households whose head or its spouse is a married woman aged 15-35 in the first wave. However, results are very similar to those presented here.

In the sequel we contrast methods relying on the Unconfoundedness Assumption (UNC), such as regressions and propensity score matching, with methods allowing for selection on unobservables, such as the Instrumental Variable (IV) estimators. We already stressed the fact that these methods are not equivalent in what they estimate. With Regressions and Propensity Score Matching (PSM) we can identify and estimate the Average Treatment Effect (ATE) and the Average Treatment effect on the Treated (ATT), while IV methods give the Local Average Treatment Effect (LATE), unless we are willing to impose additional strong assumptions.

Since LATE is the average causal effect of the treatment on the sub-group of compliers, it is generally different from ATE and ATT. Moreover, different instruments identify the effect on different groups of compliers giving different estimates of the LATE. A problem for policy making is that the compliers are, in general, an unobserved sub-group. However, IV methods estimate relevant policy parameter if the instrument itself is a potential policy variable. We explore these issues with an application on data derived from the VLSMS.

5.2 Regression and propensity score matching results

Using the terminology introduced in chapter 2, we assume in this section that the unknown assignment mechanism in action in our observational study is a regular one. That is, we assume that all the relevant variable that influence both the selection in the treatment and the outcome are observed. Therefore, in this section we present the results of the estimation of the causal effect of childbearing on consumption expenditures obtained by using multiple regression models and the propensity score matching procedure, both relying on the UNC assumption. Matching is based on the nearest neighbor method with replacement using the *nnmatch* module in *STATA* (Abadie et al, 2004)[16].

[16] This software implements the estimators suggested by (Abadie and Imbens, 2002), allowing to obtain analytical standard errors which are robust to potential heteroschedasticity. We preferred

Our choice of covariates is based mainly on dimensions which are important for both households' standard of living and fertility behaviour, and hence these variables are potentially confounders that have to be included in the conditioning set to make the UNC more plausible. More specifically, we think that all these variables can theoretically have an impact both on consumption expenditures growth and on the decision of have children. In the selection of these variables we took into account the literature on the relationship between fertility and poverty, which we discussed in section 1.4.

The covariates we use are all measured at the first wave and hence can be viewed properly as pre-treatment variables. This is an important aspect, ensuring that their values are not influenced by the treatment[17]. Therefore, the availability of data on two time points allowed us to implement properly a pre-post treatment study. Moreover, since the outcome is the difference between the consumption expenditures in the second wave (post-treatment value) and the corresponding value at the first wave (pre-treatment) we are able to control for that part of unobserved heterogeneity which is time-constant. In other words, we combine methods relying on the UNC with a difference-in-difference estimator. As mentioned in section 2.3, this strategy is expected to increase the robustness of estimates to potential unobserved confounders.

Among the covariates we included demographic characteristics of the household, such the sex and the age of the household head, the household size and the presence of children. The effect of children is distinguished by their age distribution, and expressed as a ratio of the total number of household members. Other covariates are the ratio of male and female members aged 15-45, the ratio of male and female working members aged 15-45 out of the respective groups,

---

analytical to bootstrapped standard error because Abadie and Imbens (2004) have showed that bootstrap fails with nearest neighbor matching. This matching method, on the other hand, should be preferred since, ensuring the "best" matches, it reduces the bias with respect to other methods. We compare this matching method with alternative ones in the section 5.3.3.

[17] Variables measured at the first wave could be still influenced by the treatment through some kind of "anticipation effect". For example, if a couple plans to have a child and suspects that this choice will affect negatively its wellbeing they could choose, for example, to increase their labour supply, or other behavioural parameters. It follows that the vector of covariates can also include lagged outcomes. In our application we include in the conditioning set consumption expenditure measured at the first wave. This reflects the households' level of living standard prior to treatment, and is likely to be of relevance.

an educational index, the level of equivalized consumption at the first wave and regional dummies. Importantly, we included two binary variables indicating, respectively, if the household is farmer or not and if the household head belong to the majority ethnic group (the Kinh) or not. As mentioned in section 1.5, the data also includes rich information on the characteristics of the community where the household resides. We control for community differences through three indexes: 1) an index of economic development, 2) health facilities and 3) educational infrastructures. The exact definition of these indexes, as well as that of the other variables in explained in the appendix to this chapter.

The results are presented in Table 5.2. We also report the results of the estimation of a simple regression of Y on D without any covariates. This is equivalent to use the naïve estimator (2.26) defined in chapter 2 and can be obtained also from the Table 5.1. In fact, this estimate (-462) is equal to the difference in the consumption expenditure growth between households with at least one new child (1004) and with no new child (1466; see table 5.1).

This would be an acceptable estimate of the ATE under the randomization of D. It is clear that selection is present and the estimate of fertility on expenditure is reduced by around 10% in the multiple regression. The other methods used maintain the assumption that the treatment can be thought of as randomized after having controlled for covariates (UNC). What differ are the assumptions imposed for estimation. The standard multiple regression implicitly assumes that the effect of childbearing on poverty is constant, while FILM, including all interactions among D and covariates, allows it to change with covariate values. As a consequence, multiple regression does not distinguish between ATE and ATT since they coincide under a constant treatment effect. In contrast, FILM does, and ATE and ATT will in general differ. FILM was implemented with and without conditioning on the common support. Since results are very similar we show only FILM on the common support. FILM in this case requires a first stage estimation of the propensity score and the common support[18].

[18] We used a logit specification including some interaction terms to achieve balancing. We discuss the assessment of balancing quality in section 5.3.1.

Table 5.2 - Estimates from methods based on the Uncounfoundedness Assumption (robust standard error in parentheses)

REGRESSIONS				PROPENSITY SCORE MATCHING	
SIMPLE	MULTIPLE WITH NO INTERACTIONS	FILM (conditioned on CS)			
(ATE=ATT)	(ATE=ATT)	ATE	ATT	ATE	ATT
-462	-414	-421	-432	-411	-356
(56)	(62)	(60)	(59)	(87)	(116)

Notes: CS = common support. Figures are in thousands of dongs. Standard errors for regressions are robust to heteroschedasticity and within community dependency. PSM standard errors are robust to heteroschedasticity. The matching method used is the nearest neighbor with replacement.

With FILM, multiple regression model is made as similar to PSM as possible. The difference of course, is that PSM does not impose any functional form for the relationship between poverty and fertility. Regression, in contrast imposes linearity.

As we can see from Table 5.2 the estimate for ATE is similar in all these methods, while ATT is estimated lower in the PSM. Thus, relaxing linearity matters and the PSM is to be preferred. Moreover, PSM permits us to assess, in a simple way, the underlying process of comparison between treated and control units. We assess the balancing achieved in the covariates by using the PSM in section 5.3.1.

In general, ATT and ATE differ if the distribution of covariates in the two groups of treated and control are different (this is expected due to likely potential self selection into treatment) and if the treatment interacts with covariates (treatment effect heterogeneity). Hence, we found evidence for the presence of some form of selection that covariates control for. However, if the selection was also on unobservables this biased our estimates. We address this issue in the section 5.3.4 through a sensitivity analysis.

In order to assess the magnitude of the estimate on consumption expenditure, we compare, in table 5.3, the effect of the treatment variable *D* with those of the other covariates included in the multiple regression model.

Table 5.3 – Parameter estimates of the linear model for consumption growth, *Y,* and the logit model for selection into treatment.

Variables	Outcome equation (linear regression)			Selection equation (logistic regression)		
	Coef.	Robust Std. Err.	ey/ex (elasticity)	Coef.	Robust Std. Err.	ey/ex (elasticity)
D	-413.58	66.07	-0.12			
Sexhhh	96.83	91.40	0.06	-0.02	0.18	-0.01
Kinh	173.39	61.71	0.11	-0.03	0.19	-0.02
perkids_04	-2.61	3.64	-0.03	0.01	0.01	0.14
perkids_59	-1.25	2.87	-0.02	-0.02	0.01	-0.28
perkids_1014	-8.45	2.95	-0.08	-0.03	0.02	-0.28
permale_1545	-0.73	3.74	-0.01	0.02	0.01	0.25
perfema~1545	-1.97	4.33	-0.04	0.02	0.01	0.28
Farm	-35.95	64.44	-0.02	0.11	0.13	0.05
Edu	10.00	1.63	0.42	-0.02	0.01	-0.62
rlpcex1	-0.12	0.08	-0.11	0.00	0.01	-0.01
region1	45.44	98.58	0.01	-0.03	0.24	0.00
region2	26.95	102.00	0.01	-0.49	0.25	-0.08
region3	-50.55	109.75	-0.01	0.56	0.25	0.05
region4	208.02	104.51	0.02	0.64	0.24	0.04
region5	400.58	147.09	0.01	1.61	0.37	0.04
region6	1163.51	198.20	0.07	0.30	0.27	0.01
Agehhh	4.22	4.22	0.13	0.00	0.01	-0.13
Hhsize	-0.65	16.13	0.00	-0.09	0.04	-0.35
peractm_1545	4.32	1.50	0.31	0.01	0.02	0.31
peractf_1545	0.55	1.42	0.04	0.00	0.01	0.22
IEI	12.81	15.22	0.05	-0.06	0.03	-0.20
EDI	-64.92	52.35	-0.14	0.14	0.13	0.27
HFI	22.52	18.80	0.11	-0.06	0.04	-0.23
Constant	1079.89	522.43		0.13	1.00	

The treatment variable *D* has, out of the various covariates, the third strongest elasticity. Moreover, given that the average consumption growth between the two waves amounts to 1285 thousand dongs, the estimates ranging from -356 to -432 thousand dongs, as presented in Table 5.2, are clearly substantial. In addition, the food poverty line in 1992 was estimated to 750 thousands of dongs (corresponding to 68 US$ in year 1992), which is another indication that the

effects associated with childbearing events are important for households' economic wellbeing[19].

Finally, we can compare the magnitude of the estimated effects with the cost of a fundamental good for Vietnamese: the rice. In 1992, the amount needed for buying a quantity of rice giving 1000 calories (about 300 gr.) each day for one year was equal to 215 thousand of dongs[20]. From these figures, we conclude that the estimated causal effect is not only statistically significant, but also economically relevant.

An interesting, but not unexpected, result is that education seems to be the most important confounder in the relationship between fertility and poverty. In fact, it has the strongest effect in both models. This result is in line with the theoretical literature we discussed in chapter 1.

It is interesting to note that the consumption level measured in 1992 has little impact on the probability of childbearing events taking place between the two waves, and suggest that the issue of reversed causality seems to be not relevant in our application[21].

The previous estimate of the ATT refers to the whole population of treated. However, for policy makers it would be of interest to assess if, and how, the treatment has different effects according to the specific characteristics of the treated households. Therefore, we implemented an analysis of the treatment effect heterogeneity, which is presented in Table 5.4. Bearing in mind that the overall ATT is -356 with a standard error of 116, it is clear that there is substantial variation in the treatment effect for different groups[22]. First of all, we note huge variations by regions. However, in some regions we have very few matched units hampering reliable causal effect estimation.

---

[19] For insights on the goods composing the Vietnamese food basket and for details about the construction of the Vietnamese food poverty line see Tung (2004).

[20] These figures are derived by Molini (2006).

[21] In a preliminary phase, we deepened the issue of reverse causality by using a simultaneous equations approach. We estimated simultaneously two equations: one for fertility and one for consumption. In the fertility equation we found a non-significant effect of the consumption measured at the first wave, confirming the result in table 5.3.

[22] This is confirmed by the FILM model. After we ran it, we tested for the presence of heterogeneous effects of the treatment, that is, we tested the joint significance of all the interactions between *D* and covariates. The null hypothesis of no significant interaction was rejected (F = 1.94; Prob>F = 0.0120).

Table 5.4 – Treatment effect heterogeneity

Sub-sample	Treated units			Untreated units			TOT units	ATT
	MA	UN	TOT	MA	UN	TOT		
Regional variability								
Region1	90	29	119	146	43	189	308	-433 (168)
Region2	54	31	85	230	54	284	369	-52 (232)
Region3	61	39	100	79	34	113	213	-512 (362)
Region4	45	28	73	44	41	85	158	-1093 (381)
Region5	2	35	37	1	18	19	56	-292 (na)
Region6	11	34	45	17	53	70	115	-2239 (1384)
Region7	90	17	107	186	22	208	315	-464 (212)
Farmer / non farmer								
Farmer	372	22	394	500	37	537	931	-720 (152)
Non Farmer	153	22	175	387	44	431	606	-339 (204)
Kinh / non kinh								
Kinh	409	33	442	796	24	820	1262	-328 (138)
Non Kinh	96	31	127	87	61	148	275	-407 (192)
Household Education Index								
EDU – L	228	9	237	257	16	273	510	-355 (167)
EDU – M	145	33	178	301	26	327	505	-334 (145)
EDU – H	125	29	154	323	45	368	522	-497 (257)
Consumption in 1992								
Consumption in W1 – L	158	33	191	258	63	321	512	-278 (122)
Consumption in W1 – M	174	16	190	307	16	323	513	-317 (158)
Consumption in W1 – H	164	24	188	276	48	324	512	-451 (286)
Household size								
Household size – L	189	63	252	184	49	233	485	-408(164)
Household size – M	160	25	185	426	36	462	647	-333(175)
Household size – H	108	24	132	240	33	273	405	-377 (234)
% of kids in 1992								
% kids in wave 1 – L	138	21	159	389	64	453	612	-365 (172)
% kids in wave 1 – M	116	27	143	266	29	295	438	-368 (198)
% kids in wave 1 – H	215	52	267	193	27	220	487	-364 (223)
Age of household head								
Household head age – L	181	58	239	203	35	238	477	-169 (162)
Household head age – M	125	13	138	402	20	422	560	-294(205)
Household head age – H	188	4	192	253	55	308	500	-445(284)

Notes: Estimates of the ATT are based on the PSM method (using the nearest neighbor with replacement) implemented on different sub-samples. Standard errors are in parentheses. EDU is an educational index aggregating years of schooling for all household members and keeping into account the age; L, M, H attached to same variables means "low", "medium" and "high" groups obtained by splitting the sample in three parts of equal size; MA = matched observations; UN = unmatched observations; na = cannot be estimated; W1 = first wave.

Interestingly, the distinction between farmers and non-farmer households give rise to clear heterogeneity. In particular, farmer households with an additional child are substantially more disadvantaged than non-farmers. This is also the case for non kinh versus kinh households but here the heterogeneity is smaller. The heterogeneity by education confirms the well known pattern. More highly educated individuals suffer more from a childbearing event, mainly due to differences in opportunity cost. That is, more highly educated women earn more and, consequently, suffer more from retracting from the labor market due to childbearing.

We also find that households headed by older individuals suffer more from childbearing events. Interestingly, the percentage of children in the first wave seems less important, indicating that economies of scale are likely to be not so relevant. Finally, the effect of the household size forms an *U*-shape: the (negative) effect of childbearing is stronger for small and large households while medium sized households show a somewhat lower effect.

The rather strong heterogeneity in the treatment effects suggests, of course, that care is needed when implementing policy interventions. Clearly, policies related to fertility and economic wellbeing will have very different effects for different groups. Moreover, as Crump et al (2006) notice, if there is strong evidence in favor of heterogeneous effects, one may be more reluctant to recommend extending the program to populations with different distributions of the covariates. This confirm the idea that the fertility-poverty relationship is country-specific and should be assessed case by case, without any possibility of generalize the results found in a specific country.

5.3 Assessing the PSM procedure

In this section we try to assess the robustness of the estimates we get with the PSM procedure, which rely on several assumptions and technical choices. As discussed in chapter 2, the fundamental identifying assumption is the unconfoundedness, which is not directly informed by the data. We already

noticed that the reasonability of this assumption should be evaluated by the scientist's knowledge about the studied phenomena. However, even if UNC is not formally testable, we can assess the consequence of its violation through a sensitivity analysis. Some indirect tests were also proposed in the literature.

Another aspect of the PSM procedure is that its implementation requires several choices, including the matching method to use and the way to determine the Common Support (CS). After the matching algorithm has been employed, we have to evaluate the quality of balancing we get in the covariates. We explore these issues in the sequel of the section. We will also address the robustness of PSM estimates to the specification of equivalence scale. This last aspect is more concerned with the specificity of the application at end than with the PSM procedure *per se.*

5.3.1 Covariate balancing after matching

Since we do not condition on all covariates but on the propensity score, it has to be checked if the matching procedure is able to balance the distribution of the relevant variables in both the control and treatment group. Several procedures to do so are available (see for details Caliendo and Kopeining, 2005). These procedures can also help in determining which interactions and higher order terms to include in the propensity score specification for a given set of covariates. The basic idea of all approaches is to compare the situation before and after matching and check if there remain any differences after conditioning on the propensity score. If there are differences, matching was not (completely) successful and remedial measures have to be done, e.g. by including interaction-terms in the estimation of the propensity score. A helpful theorem in this context was presented in chapter 2. We referred to it as the balancing property of the propensity score (theorem 2.1, pag. 53). We remember that this theorem means that after conditioning on the propensity score, additional conditioning on covariates should not provide new information about the treatment decision. Hence, if after conditioning on the propensity score there is still dependence on X, this suggests either mis-specification in the model used to estimate $P(D = 1|X)$

(Smith and Todd, 2005) or a fundamental lack of comparability between both groups (Blundell et al, 2005). In our work, we used two methods to assess the balancing property: the stratification test and the comparison between standardised absolute bias before and after matching.

The stratification test was proposed by Dehejia and Wahba (1999), and starts by splitting observations into strata based on the estimated propensity score, such that no statistically significant difference between the mean of the estimated propensity score in both treatment and control group remain. Then, they propose to use t-tests within each stratum to test if the distribution of covariates is the same between both groups. If there are remaining differences, they suggest to add higher-order and interaction terms in the propensity score specification, until such differences no longer emerge. This procedure is implemented by the *pscore* procedure in *STATA* (Becker and Ichino, 2002).

We used this procedure as a support in our model building, since it suggests the interactions and higher order terms to be included in the propensity score to improve the balancing. Following this procedure we added some interaction terms (7) but no higher order terms[23]. The final number of strata was 10, ensuring that the mean propensity score was not different for treated and controls within each stratum. Moreover, the final specification balanced the covariates within these blocks. However, as remarked by Lee (2006) balancing tests have to be specific to the matching method used since each method conduct to a different control group. Therefore, the stratification method is a valid balancing test procedure only when we use stratification also as the matching method. For the other matching methods that we used we assessed the balancing quality by calculating the absolute standardized bias (ASB), before and after matching.

The ASB, suggested by Rosenbaum and Rubin (1985), is defined as the absolute difference of sample means in the treated and matched control

[23] We avoided the inclusion of higher order terms because, as demonstrated by Zhao (2005) their inclusion could have some biasing effect (while he found that the inclusion of interactions has not this drawback). Note that, given the purpose of balancing relevant observed covariates, the estimation of the propensity scores does not need a behavioral interpretation.

subsamples as a percentage of the square root of the average of sample variances in both groups. In formula, the ASB is given by:

$$ASB = \left| 100 \frac{\left(\overline{X}_T - \overline{X}_C\right)}{\sqrt{0.5\left(s_T^2 - s_C^2\right)}} \right| \quad , \tag{5.1}$$

where for each covariate $\overline{X}_T$ and $\overline{X}_C$ are the sample means, respectively, in the treated and control group and s_T^2 and s_C^2 are the corresponding sample variances. One possible problem with the standardised bias approach is that one does not have a clear indication for the success of the matching procedure, even though in most empirical studies a standardised bias below 3% or 5% after matching is seen as sufficient.

In table 5.5 we present, for each covariate included in the final specification of the propensity score, the ASB before and after matching. We note that exists considerable initial bias between households who experience at least one childbearing event and households who did not. For instance, ten covariates have initial standardized differences larger than 20%. Particularly, high are the standardized biases shown by several demographic characteristics of the household (e.g., the percentage of kids, in all the three groups of age, show bias larger than 50%). From table 5.5 we note that matching performed well in reducing the bias of background variables. For instance, the initial absolute standardized bias for "Edu" (the educational index) was initially 36.54% and the matching reduces it to 3.50%. Importantly, matching reduced the standardized bias for each covariate, with the irrelevant exception of "peractf_1545" (percentage of female members aged 15-45). For this variable, the bias after matching is, however, 5.06%.

Table 5.5 – Absolute standardised bias before and after matching for each covariate included in the propensity score specification.

Variable	Sample	Treated	Control	ASB (%)
Sexhhh	Before	0.845	0.880	9.75
	After	0.847	0.857	2.93
Kinh	Before	0.777	0.847	18.18
	After	0.776	0.779	0.93
perkids_04	Before	19.981	12.492	50.05
	After	19.926	19.841	0.06
perkids_59	Before	11.538	19.871	54.30
	After	11.663	11.827	1.12
perkids_1014	Before	7.464	15.573	60.31
	After	7.542	7.550	0.10
permale_1545	Before	25.484	21.254	41.04
	After	25.119	24.985	1.03
perfema~1545	Before	26.127	23.057	34.04
	After	25.766	25.697	0.08
farm	Before	0.692	0.555	28.07
	After	0.687	0.679	1.05
edu	Before	49.030	57.568	36.54
	After	49.001	49.824	3.50
rlpcex1	Before	1065.025	1116.021	8.50
	After	1067.094	1046.090	3.05
region1	Before	0.209	0.195	3.05
	After	0.217	0.209	1.08
region2	Before	0.149	0.293	35.20
	After	0.154	0.137	4.05
region3	Before	0.175	0.116	16.07
	After	0.167	0.164	0.80
region4	Before	0.128	0.088	13.01
	After	0.131	0.126	1.08
region5	Before	0.070	0.019	24.06
	After	0.066	0.056	4.04
region6	Before	0.079	0.072	2.06
	After	0.080	0.078	0.07
agehhh	Before	39.874	40.571	5.23
	After	39.718	40.227	2.82
hhsize	Before	5.223	5.772	26.07
	After	5.278	5.311	2.54
peractm_1545	Before	95.357	91.253	17.08
	After	95.279	96.357	4.73
peractf_1545	Before	93.052	94.267	4.62
	After	93.103	94.414	5.07
IEI	Before	5.182	5.674	20.84
	After	5.220	5.287	2.96
EDI	Before	2.761	2.794	6.02
	After	2.765	2.792	0.92
HFI	Before	6.080	6.301	11.76
	After	6.087	6.151	3.49

These results suggest that the balancing property is satisfied, meaning that matched treated and control can be considered sufficiently similar (on average)[24]. If we refer to the bias decomposition discussed in section 2.3 (pag. 45), we can say that the PSM procedure has virtually cancelled out the bias term B_1. In the next sections we assess the potential relevance of the other two sources of bias.

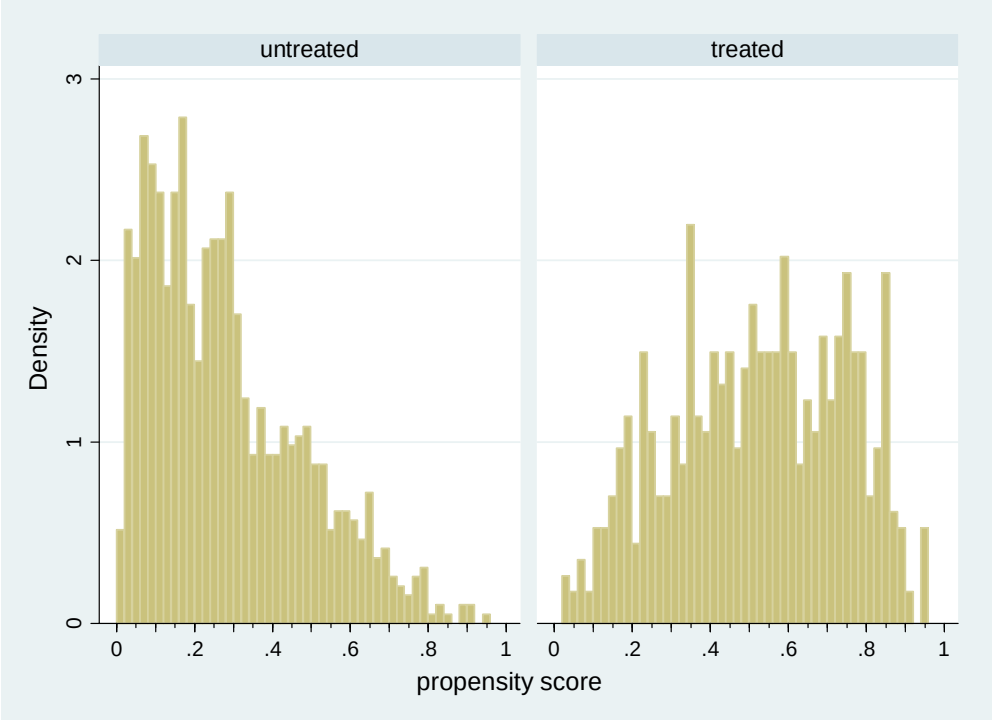
5.3.2 Evaluating the overlap

In chapter 2 we highlighted that ATT and ATE are only defined in the region of common support. Heckman et al (1997) point out, that a violation of the common support condition is a potentially relevant source of evaluation bias. Comparing the incomparable must be avoided, i.e. only the subset of the comparison group that is comparable to the treatment group should be used in the analysis (Dehejia and Wahba, 1999). To this end, we remember that our previous analyses were implemented on the sub-sample of household with at least one married woman aged 15-40 in the first wave. As we already said, this selection is part of the whole matching strategy, since it allow to exclude from controls those households who are "ineligible" because of absence of fecund woman.

However, an important step is to check the overlap and the region of common support between treatment and comparison group in the selected sub-sample. Several ways are suggested in the literature, where the most straightforward one is a visual analysis of the density distribution of the propensity score in both groups, as presented in Figure 5.1. This figure indicates an expected pattern, where in the left tail of the propensity score distribution there is a prevalence of untreated, while approaching the right tail untreated density becomes lower and lower. From figure 5.1 we get an indication that treated and controls overlap sufficiently. In fact, there are no large intervals where units of only one group are missing.

[24] Similar good results were obtained also using different matching methods.

Figure 5.1 – Distribution of the propensity score in the treated and control group



However, several methods can be used to determine the region of common support more precisely[25]. The simplest is, essentially, based on comparing the minimum and maximum of the propensity score in both groups (min-max criterion). In our case, the estimated propensity score lies between 0.00486 and 0.94315 for control and between 0.02061 and 0.95379 for treated. Hence, the common support, determined by the min-max criterion, is 0.02061, 0.94315. As noticed by Becker and Ichino (2002), limiting estimates on the common support could improve matching quality, but this could not be the case when dropping observation out of the support we exclude some good match for

[25] These methods are reviewed by Crump et al (2007). They also suggest that the limited overlap problem can be also addressed by changing the estimand of interest. This is a method to assess robustness with respect to possible common support failure. Lechner (2000) argues that either ignoring the common support problem and estimate causal effects only for subpopulation on the common support could be misleading. Ignoring the problem may result in bias due to non comparable control and treatment groups. Discarding observations out of the common support could give inconsistent estimates especially in the case of heterogeneous treatment effect. He suggests a method to bounding the treatment effects in case of failure of the CS condition.

observations that are close to the bounds of the support. However, in our case we have only 16 units (16/1537 = 1%) out of Common Support (10 controls (10/968 =1%) and 6 (6/569 = 1%) treated). These 16 units are the only ones that we discard using some matching methods (kernel and nearest neighbour with replacement). As noticed by Bryson et al (2002), when the proportion of lost observations is small, discarding observations out of the common support poses few problems. Therefore CS seems not to be a problematic issue in our application.

Nevertheless we note that the figure 5.1 shows, as usual, quite low densities in the tails of the propensity score distribution. Therefore, we explore an alternative method to determine the common support: the trimming procedure, suggested by Heckman et al (1997).

The trimming method consists to drop observations which estimated propensity score density is below a certain threshold. We use three different thresholds: 2% (as used, for example, by Heckman et al, 1997; Smith and Todd, 2005), 3%, 5%.

In the table 5.6 we show a sensitivity analysis to different methods to determine the CS. The matching method is still the nearest neighbour with replacement. As we can see from the table, the estimated ATE and ATT are quite robust and the number of discarded units is limited both using the min-max criterion and trimming method. In the table we included also the estimate restricted to the so-called "thick support", which is a way to assess robustness of estimates with respect to subject in the tails (Black and Smith, 2004). The "thick support" is defined as the region so that $0.33<P(D=1|X)<0.67$. From table 5.6 we can see that estimates based on the "thick support" are very similar to the baseline ones. This can be taken, as suggested by Black and Smith (2004) and Eren (2006), as evidence that measurement error and selection on unobservable have at most a modest effect on the estimated causal effect. However, imposing the "thick" support drastically reduces the sample size and, as a consequence, the ATT become statistically non-significant.

Table 5.6 – Sensitivity analysis to different ways to determine the common support (Standard error in parentheses).

Common support method	Treated			Untreated			ATT	ATE
	M	U	TOT	M	U	TOT		
None	554	0	569	964	0	968	-353	-411
							(129)	(86)
Minima-maxima	549	20	569	957	11	968	-356	-411
							(116)	(87)
Trimming 2%	544	25	569	946	22	968	-337	-397
							(113)	(84)
Trimming 3%	540	29	569	934	34	968	-324	-383
							(107)	(76)
Trimming 5%	540	29	569	915	53	968	-318	-371
							(119)	(79)
Thick support (0.33 – 0.67)	266	303	569	268	700	968	-355	-432
							(209)	(150)

Note: M = matched; U = unmatched (discarded units).

5.3.3 Sensitivity to the matching algorithm

As briefly discussed in chapter 2, several methods can be used to match control and treated units. We already noted that, asymptotically, all PSM estimators should yield the same results, while in small samples the choice of the matching algorithm can be important, where usually a trade-off between bias and variance arises. It is, also, clear that there is no "winner" for all situations and that the choice of the estimator crucially depends on the situation at hand.

We explored this issue by comparing estimates from different methods. These are presented in table 5.7, along with the units discarded by the several procedures. Since the estimates are robust to the matching method, the choice is not crucial. We note that by using some methods (kernel methods and the nearest neighbour with replacement and without caliper) we discard only units out of the common support, which in the table 5.7 is determined by the min-max criterion. In the following analyses, likewise those presented in the previous sections, we use the nearest neighbour method with replacement (and without caliper).

Table 5.7 – Sensitivity to the matching method (standard errors in parentheses)

Method	ATT	ATE	Treated			Untreated		
			M	U	TOT	M	U	TOT
nn with replacement and caliper(0.01)	-361 (118)	-414 (83)	549	20	569	957	11	968
nn without replacement; with caliper (0.01)	-429 (87)	-419 (81)	397	172	569	405	563	968
nn with replacement (without caliper)	-356 (116)	-411 (87)	563	6	569	958	10	968
nn without replacement (without caliper)	-414 (67)	-454 (59)	563	6	569	563	405	968
radius matching (0.01)	-406 (88)	-416 (74)	549	20	569	957	11	968
gaussian kernel	-426 (79)	-416 (69)	563	6	569	958	10	968
epanechnikov kernel	-446 (86)	-423 (71)	563	6	569	958	10	968

Note: nn = nearest neighbour; M = matched; U = unmatched (discarded units); when a caliper is impose to the nn or the radius method is used the tolerance level was set to 0.01 (that is the propensity score for matched treated and control should differ at most for 0.01).

5.3.4 Assessing the unconfoundedness assumption

In the previous sections we assessed the matching quality, the overlap and the robustness of the PSM to the matching method. However, the most critical requirement of the PSM is that UNC holds, which is not directly testable by the data. It is therefore of interest to assess the extent parameter estimates might be affected by any violation of the UNC.

Several approaches are proposed in the literature (see Ichino et al, 2007 or Imbens, 2004 for an extended discussion). A first strategy is to focus on estimating the causal effect of a treatment that is known to have a zero effect, e.g. by relying on the presence of multiple control groups, as suggested by Rosenbaum (1987). If one has a group of eligible and ineligible non-participants, the treatment effect which is known to be zero can be estimated using only the two control groups (where the treatment indicator then has to be a dummy for belonging in one of the two groups).

Any non-zero effect implies that at least one of the control groups is invalid. In our application a second control group is given by the households with no woman in the fecund age at the first wave ("ineligible" to have children).

Another idea is to estimate the causal effect of the treatment on variables known to be unaffected by it, typically because their values are determined prior to the treatment itself (Imbens, 2004). If this is not zero, this implies that the treated observations are distinct from the controls; otherwise it is more plausible that the unconfoundedness assumption holds. If the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. In our context, we can think to measure the effect of the treatment *D,* as defined in section 5.1, on covariates measured at the first wave. We expect that having a child between 1992 and 1997 had no impact on variables measured in 1992[26].

We present a set of indirect tests of the UNC in the table 5.8. The estimated ATT are very small, as required, and in any case not significant. These indirect tests, however, are far from being definitive about the validity of the UNC.

An alternative approach is to implement a sensitivity analysis in order to assess the effect on the estimates of a departure from the UNC. We adopt the simulation-based approach suggested by Ichino, Mealli and Nannicini (2007 – in the following IMN). The underlying hypothesis is that assignment to treatment may be confounded given only the set of observables covariates X (i.e. the UNC does not hold) but it is unconfounded given X *and* an unobservable covariate U[27]. Thus, $Y_0 \perp D \mid (X, U)$. By changing the assumptions about the distribution of U, we can assess the robustness of the ATT with respect to different hypotheses regarding the nature of the confounding factor. Moreover, we can verify whether there exist a set of plausible assumptions on U under which the estimated ATT is driven to zero, or very far away, by the inclusion of U in the matching set.

[26] However, if the decision to have a child was taken prior to the first wave there could be some sort of "anticipation effects". This is unlikely to hold in the context of our application.

[27] This assumption is used also in other works (e.g. Imbens, 2003), but the approach used here has the advantage of allowing to assess point estimates sensitivity without relying on parametric estimation of the outcome.

Table 5.8: Indirect tests of unconfoundedness

Test	ATT	Standard error
Using a second control group	14.08	56.97
Using the following variables as outcome:		
Consumption in 1992	-32.95	42.60
Sex of household head in 1992	0.02	0.03
Age of household head in 1992	-1.51	0.95
Kinh indicator in 1992	-0.01	0.02
%kids (0-4) in 1992	-0.86	1.53
%kids (5-9) in 1992	-1.33	1.29
Farm indicator in 1992	0.05	0.04
Educational index in 1992	-1.49	2.07
Household size in 1992	-0.21	0.18

Note: the first test compares two control groups. The first is the one used in the previous analyses: households with at least one married woman aged [15-40] at the first wave and with no childbearing events between the two waves. The second is represented by household with no woman aged [15-45] at the first wave. The "treatment" in this case consist to belong to the first control group. The other tests, estimate the effect of childbearing events on the specified outcome.

IMN assume in the simulation of the *U*-values that also the outcome is a binary variable. If the outcome is continuous, which is our case, a transformation is needed so that the outcome takes the value 1 if it is above a certain threshold (the median for example) and 0 otherwise, alternatively one could consider other outcome variables such as poverty status which essentially is a dichotomous transformation of consumption expenditure[28]. However, in the final step of the simulation we estimate the effect on the original specification of the outcome, which in our case is a continuous one.

The potential confounder can be specified in different ways. One alternative is a "calibrated" version, where we make the distribution of *U* similar to the empirical distribution of important binary covariates in the set *X*. Another

[28] For technical details on the simulations, see Ichino et al (2007) and Nannicini (2007) for details on the *STATA* module *sensatt*, which implements the sensitivity analysis.

alternative is to specify a "killer" confounder, where the values of U are specified so that its association with the outcome and the treatment is increasingly high. In this sensitivity analysis, we are particularly worried about the fact that the estimated ATT could become zero, or non significant, when including U. This is an interesting approach, because it gives us a measure of how large the association between U, Y_0 and D has to be in order to cancel out the ATT. The distribution of U is specified by the following four key parameters:

$$p_{ij} = P\,(U=1|D=i,\ Y=j\,) = P\,(U=1|D=i,\ Y=j\,,\ X) \quad i,j = 0,1 \qquad (5.2)$$

In (5.2) the hypothesis (used in the simulation) that U is independent to X conditional to D and Y is made[29]. In order to choose the signs of the associations, IMN note that if $d = p_{01} - p_{00} > 0$, then U has a positive effect on Y_0 (conditioning on X) whereas if $s = p_{1.} - p_{0.} > 0$, where $p_{i.} = P(U=1|D=i)$, then U has a positive effect on D. If we set $p_u = P(U=1)$ and $d' = p_{11} - p_{10}$, then the four parameters p_{ij} are univocally identified, specifying the values of d and s. Hence, we can fix the two quantities p_u and d', which will not affect the baseline estimate, and change the values of d and s to assess the sensitivity of the estimates.

Table 5.9 shows the sensitivity analysis to some calibrated confounders, while the results for the "killer" confounder are given in Tables 5.10a - 5.10d. In these tables, the quantities G and A are defined as:

$$G = \frac{P(Y=1|\ D=0\ ,\ U=1\ ,\ X) \,/\, P(Y=0|\ D=0\ ,\ U=1\ ,\ X)}{P(Y=1|\ D=0\ ,\ U=0\ ,\ X) \,/\, P(Y=0|\ D=0\ ,\ U=0\ ,\ X)} \qquad (5.3)$$

and

[29] The authors, in a previous version of the work (Ichino et al, 1996) assessed the relevance of the assumptions used in the simulation through Monte Carlo studies. They found that imposing a binary confounder, instead of a continuous one, and the independence of U and D do not sensibly affect the validity of the analyses.

$$A = \frac{P(D=1|U=1, X) / P(D=0|U=1, X)}{P(D=1|U=0, X) / P(D=0|U=0, X)} \quad . \qquad (5.4)$$

The parameter *G,* in the (5.3), is the average odds ratio from the logit model of $P(Y=1| D=0, U, X)$ calculated over 1000 iterations. It is, in words, a measure of the effect of *U* on *Y*, and we refer to as the *outcome effect*. The parameter *A*, in the (5.4), refers to the average odds ratio from the logit model of $P(D=1|U, X)$. This is a measure of the effect of *U* on *D*, and is therefore a measure of the *selection effect*. Therefore, these parameters are important in order to evaluate the magnitude of the associations that makes the ATT sensibly different from the baseline estimate.

The first simulation in table 5.9 sets the distribution of *U* to be similar to the distribution of the household head sex. In this case, given that 82% of the households who were exposed to treatment (*D*=1) and showed a consumption growth higher than the median (*Y*=1) have a male head, by setting $p_{11} = 0.82$ we impose that an identical fraction of households are assigned a value of $U|(T=1, Y=1)$ equal to 1. An analogous interpretation holds for the other probabilities p_{ij}. When controlling for observables, *U* has a slightly negative effect both on the outcome ($G = 0.8 < 1$) and on probability of being treated ($A = 0.92 < 1$). Under a deviation from the UNC with these characteristics, the ATT is estimated to be equal to -406. This estimate is qualitatively in line with the one obtained assuming no confounding effects, and remains statistically significant[30]. Similar results hold in the other simulations.

[30] In order to compute a standard error of the ATT estimator when *U* is included in the set of matching variables, IMN considered the problem of the unobserved confounding factor as a problem of missing data that can be solved by multiply imputing the missing values of *U.* See Ichino et al (2006) for details.

Table 5.9 – Sensitivity analysis to "calibrated" unobserved confounders

	Fraction $U = 1$ by treatment / outcome				Outcome effect (G)	Selection effect (A)	ATT
	p_{11}	p_{10}	p_{01}	p_{00}			
No unobserved confounders	---	---	---	---	---	---	-356 (111)
Unobserved confounder similar to:							
Sex of hhh	0.82	0.86	0.85	0.87	0.80	0.92	-406 (129)
Kinh	0.86	0.73	0.79	0.78	2.46	0.64	-412 (132)
% children 0-4	0.65	0.53	0.41	0.41	1.16	2.58	-440 (146)
% women 15-45	0.70	0.58	0.45	0.41	1.15	2.15	-445 (142)
Farmer	0.64	0.72	0.55	0.56	0.99	1.83	-440 (139)
EDU	0.45	0.40	0.62	0.50	1.68	0.55	-421 (135)
Consumption W1	0.41	0.36	0.42	0.37	1.24	0.93	-419 (135)
Age of hhh	0.38	0.35	0.37	0.36	1.05	1.02	-413 (134)

Note: Estimates based on the nearest neighbour matching. Standard errors are analytical and each reported ATT is the average over 1000 iterations. The parameters p_{11}, p_{10}, p_{01} and p_{00} characterise the distribution of U and are defined in the formula (5.2).

The key conclusion from table 5.9 is that for different simulated U, the resulting ATT is never substantially different from the ATT based on the UNC assumption (-356). Thus, even if an unobservable variable (with the same distribution as those listed) was excluded from the conditioning set in the PSM, the effect on the estimated ATT would be negligible.

In tables 5.10a-5.10d we go a step further, imposing "killer" confounders. That is, we explore the robustness of the ATT by choosing the parameters d and s so that the outcome and selection effects are increasingly large. We implement four separates sensitivity analyses according to the signs of the association between U and D, U and the binary transformation of Y.

Table 5.10a – Sensitivity analysis to confounders so that $G < 1$ and $A < 1$

	s = -0.1	s = -0.2	s = -0.3	s = -0.4	s = -0.5
d= -0.1	-442	-452	-457	-464	-460
	(143)	(157)	(172)	(193)	(217)
	G=0.73	*G*=0.71	*G*=0.67	*G*=0.66	*G*=0.62
	A=0.48	*A*=0.29	*A*=0.17	*A*=0.09	*A*=0.04
d= -0.2	-446	-456	-463	-466	-471
	(139)	(150)	(168)	(179)	(207)
	G=0.42	*G*=0.39	*G*=0.38	*G*=0.35	*G*=0.31
	A=0.56	*A*=0.34	*A*=0.20	*A*=0.11	*A*=0.04
d= -0.3	-451	-446	-461	-472	-477
	(135)	(146)	(161)	(173)	(196)
	G=0.25	*G*=0.22	*G*=0.20	*G*=0.18	*G*=0.14
	A=0.64	*A*=0.40	*A*=0.23	*A*=0.12	*A*=0.05
d= -0.4	-446	-441	-470	-469	-480
	(136)	(142)	(156)	(169)	(188)
	G=0.12	*G*=0.12	*G*=0.10	*G*=0.07	*G*=0.05
	A=0.74	*A*=0.45	*A*=0.26	*A*=0.13	*A*=0.04
d= -0.5	-434	-446	-471	-468	n.a.
	(135)	(140)	(152)	(161)	
	G=0.07	*G*=0.05	*G*=0.04	*G*=0.03	
	A=0.86	*A*=0.52	*A*=0.29	*A*=0.15	

Note: n.a. = combination resulting in inadmissible values of the parameters characterising the distribution of *U*.

Table 5.10b – Sensitivity analysis to confounders so that $G > 1$ and $A < 1$

	s = -0.1	s = -0.2	s = -0.3	s = -0.4	s = -0.5
d= +0.1	-445	-448	-449	-446	-435
	(150)	(168)	(195)	(217)	(251)
	G=2.03	*G*=1.93	*G*=2.08	*G*=2.17	*G*=2.14
	A=0.36	*A*=0.22	*A*=0.13	*A*=0.07	*A*=0.03
d= +0.2	-444	-445	-444	-433	-417
	(154)	(177)	(202)	(238)	(290)
	G=3.33	*G*=3.27	*G*=3.77	*G*=3.79	*G*=4.28
	A=0.31	*A*=0.18	*A*=0.11	*A*=0.06	*A*=0.02
d= +0.3	-443	-435	-437	-415	-402
	(163)	(186)	(216)	(263)	(324)
	G=5.99	*G*=6.33	*G*=6.53	*G*=7.58	*G*=9.84
	A=0.26	*A*=0.15	*A*=0.09	*A*=0.04	*A*=0.02
d= +0.4	-436	-417	-417	-383	-359
	(172)	(197)	(238)	(308)	(435)
	G=11.57	*G*=13.01	*G*=13.29	*G*=18.17	*G*=31.28
	A=0.21	*A*=0.12	*A*=0.07	*A*=0.03	*A*=0.01
d= +0.5	-420	-398	-369	-338	-238
	(184)	(219)	(271)	(388)	(639)
	G=36.02	*G*=30.67	*G*=51.94	*G*=117.46	*G*=2128.61
	A=0.17	*A*=0.09	*A*=0.05	*A*=0.02	*A*=0.01

Table 5. 10c – Sensitivity analysis to confounders so that *G* < 1 and *A* > 1

	s = +0.1	s = +0.2	s = +0.3	s = +0.4	s = +0.5
d= -0.1	-431	-440	-445	-448	-450
	(135)	(137)	(149)	(171)	(192)
	G=0.73	*G*=0.73	*G*=0.76	*G*=0.74	*G*=0.69
	A=1.19	*A*=1.85	*A*=2.94	*A*=4.79	*A*=8.43
d= -0.2	-445	-432	-448	-440	-436
	(135)	(140)	(156)	(175)	(200)
	G=0.45	*G*=0.43	*G*=0.43	*G*=0.43	*G*=0.38
	A=1.39	*A*=2.16	*A*=3.43	*A*=5.69	*A*=10.03
d= -0.3	-441	-436	-441	-437	-441
	(133)	(147)	(166)	(189)	(224)
	G=0.25	*G*=0.26	*G*=0.25	*G*=0.24	*G*=0.22
	A=1.62	*A*=2.56	*A*=4.10	*A*=6.98	*A*=12.47
d= -0.4	-432	-435	-431	-427	-415
	(136)	(149)	(172)	(201)	(247)
	G=0.14	*G*=0.15	*G*=0.14	*G*=0.13	*G*=0.11
	A=1.90	*A*=3.05	*A*=4.99	*A*=8.73	*A*=16.41
d= -0.5	-429	-430	-422	-402	-389
	(142)	(163)	(189)	(222)	(293)
	G=0.08	*G*=0.08	*G*=0.07	*G*=0.06	*G*=0.05
	A=2.27	*A*=3.74	*A*=6.39	*A*=11.47	*A*=23.49

Table 5.10d– Sensitivity analysis to confounders so that *G* > 1 and *A* > 1

	s = +0.1	s = +0.2	s = +0.3	s = +0.4	s = +0.5
d= +0.1	-428	-451	-441	-460	-459
	(136)	(138)	(143)	(155)	(173)
	G=2.12	*G*=2.12	*G*=2.16	*G*=2.18	*G*=2.40
	A=1.18	*A*=1.38	*A*=2.19	*A*=3.54	*A*=6.08
d= +0.2	-442	-434	-442	-456	-465
	(134)	(137)	(136)	(150)	(169)
	G=3.29	*G*=3.85	*G*=4.17	*G*=4.49	*G*=5.28
	A=1.16	*A*=1.18	*A*=1.88	*A*=3.08	*A*=5.33
d= +0.3	-441	-416	-447	-447	-461
	(135)	(133)	(135)	(148)	(163)
	G=6.32	*G*=7.97	*G*=7.78	*G*=8.07	*G*=10.59
	A=1.05	*A*=1.03	*A*=1.65	*A*=2.67	*A*=4.65
d= +0.4	-441	-424	-452	-449	-465
	(136)	(134)	(135)	(145)	(155)
	G=13.09	*G*=15.94	*G*=40.14	*G*=24.41	*G*=49.47
	A=1.05	*A*=1.08	*A*=1.41	*A*=2.34	*A*=4.14
d= +0.5	-433	-437	-435	-450	-465
	(142)	(134)	(137)	(142)	(152)
	G=40.72	*G*=83.74	*G*=232.02	*G*=143.41	*G*=172.65
	A=1.06	*A*=1.05	*A*=1.21	*A*=2.04	*A*=3.64

As we can see from Tables 5.10a-5.10d the estimated effect is always negative and in the majority of cases significant, as well as not dramatically different from the baseline estimate. In the first table, 5.10a, where both the outcome and selection effects are negative (odds ratios < 1) the estimates are always significant, even for very low values of *G* and *A*. The higher difference between the ATT in this table and the baseline estimate is obtained when *d* is set to -0.4 and *s* is set to -0.5. In this case the ATT is sensibly higher in absolute values (-480) than the baseline, but the outcome and selection effects are very strong (*G* = 0.05 ; *A* = 0.04).

From the other tables we note that only if the associations of *U* with *D* and/or *Y* are very strong the ATT becomes lower (in absolute value) than the baseline and not significant. This happens in particular if the effects of *U* on *D* and on *Y* have opposite signs. It is interesting to note that a theoretical relevant omitted variable such as unobserved ability, has this characteristic since it is (potentially) positively associated with *Y* and negatively associated with *D*.

For example, we get the weakest ATT (= -238) in the table 5.10b, corresponding to *d* = +0.5 and *s* = -0.5. However, in this case both the outcome and the selection effects are unreasonably strong (*G* = 2128.61; *A* = 0.01). A way to assess how *strong* the outcome and selection effects are (measured, respectively, by *G* and *A*) is to compare the odds ratios *G* and *A* with those presented in the Table 5.11, where we estimated two separate logit models, taking as outcomes *D* and the binary transformation of the outcome. For example, we can see from this table that very few covariates show odd ratios higher than 2 and lower than 0.7.

The conclusions we get from this sensitivity analysis is that ATT estimated through PSM is rather robust to the presence of potentially omitted variables. Only if the effect of this unobserved confounder (measured by *G* and *A*) would be unreasonably strong (compared to the effects of observed covariates) the estimated effect becomes insignificant.

Table 5.11 – Estimated odds ratios in the logit models for the treatment (*D*) and the binary transformation of the outcome (*Y*).

Covariates	Odds ratios	
	Y	*D*
D	0.41	---
Sexhhh	0.95	0.98
Kinh	1.71	0.97
Perkids_04	0.78	1.31
perkids_59	0.79	0.97
perkids_1014	0.97	0.96
permale_1545	0.99	1.02
perfema~1545	0.97	1.02
Farm	0.98	1.12
Edu	3.02	0.68
rlpcex1	0.99	0.99
region1	0.90	0.97
region2	0.85	0.61
region3	0.70	1.75
region4	1.43	1.89
region5	1.34	4.99
region6	5.65	1.35
Agehhh	0.99	0.99
Hhsize	0.97	0.91
peractm_1545	0.99	1.01
peractf_1545	0.99	1.00
IEI	1.57	0.95
EDI	0.91	1.16
HFI	1.72	0.95

5.3.5 Sensitivity to the equivalence scale

In this section we evaluate if the estimated effect is robust to the imposed equivalence scale. Formally, the number of adult equivalent, *Ne*, in each household is given by:

$$Ne = \left(Na + \alpha \times Nc\right)^{\theta} \quad , \tag{5.5}$$

where *Na* and *Nc* stand for the number of adult and children, respectively. The parameter α represents the adult-equivalence of a child, and the parameter θ, which is often referred to as "size elasticity", reflects possible economies of

scale. Both parameters take a value between 0 and 1. How these parameters should be calculated is still subject to debate, and there is no consensus on the matter (World Bank, 2005).

There are two possible solutions to this problem: either pick a scale that seems reasonable on the grounds that even a bad equivalence scale is better than none at all, or try to estimate a scale typically based on observed consumption behaviour from household surveys. However, there are so far no satisfactory methods for estimating the parameters of equivalence scale (Deaton and Zaidi, 2002). Moreover, it is a common empirical finding that the effect of demographic variables, such as the household size or the number of children, is weakened when imposing equivalence scales that are different from per capita expenditure (White and Masset, 2003; Balisacan *et al*, 2003).

As noted by Lanjouw and Ravallion (1995), it is always possible to impose the parameters of the equivalence scale so that the effect of demographic variables is cancelled out.

In chapter 1, we said that the equivalence scale we use is quite simple, consisting to assuming no economies of scale (θ=1) and to imposing a weight for each child (α) equal to 0.65 an adult. Given the conceptual and empirical problems involved by the formulation of equivalence scale, it is important to carry out a sensitivity analysis. This is presented in table 5.12, where can see how the ATT change by imposing different values of the parameters α and □.

From formula (5.5) is evident that the lower the values of α or □ are, the higher the values of consumption expenditures are, in both waves. Since in this period consumption growth was positive, this fact implies an increasing mean (or median) equivalent consumption growth. The ATT follows this trend in the opposite direction: the higher the consumption growth is, the lower the effect of childbearing is. This result is intuitive. In fact, the lower is α, the lower is the child-weight, and hence the childbearing effects. On the other hand, the lower is □ the higher is the size elasticity, and hence the lower is the effect of increase the household size. When these two effects are combined, we get a very weak effect of a new child entering the household.

Table 5.12 – Sensitivity analysis of the ATT estimated by PSM to different imposed equivalence scales (standard error in parenthesis).

α	□	Median equivalent consumption growth	ATT
1.00	1.00	920.89	-459
			(106)
0.80	1.00	1002.94	-408
			(97)
0.65	1.00	1073.12	-356
			(116)
0.50	1.00	1151.41	-298
			(117)
0.35	1.00	1237.30	-262
			(119)
0.65	0.80	1418.76	-306
			(160)
0.65	0.70	1641.79	-256
			(124)
0.65	0.60	1900.90	-186
			(153)
0.65	0.50	2201.90	-92
			(256)
1.00	0.50	2058.45	-227
			(228)
0.50	0.50	2276.22	-140
			(258)

We have to note, that in our sub-sample treated households have, on average, more members than controls and, as a consequence, the effect of imposing increasing economies of size is more pronounced for such households. This fact implies that consumption expenditures become more similar between treated and control households, and hence the ATT becomes lower.

From our sensitivity analysis, however, we see that the estimated effect is quite robust to several specifications of the equivalence scale. Only with "very strong" equivalence scale the ATT becomes not significant. In particular, ATT is not significant as soon as the size elasticity is set to 0.5.

5.4 Two proposed instrumental variables for the identification of the causal effect of fertility on poverty

All the previous analyses relied on the assumption that a regular assignment mechanism has regulated the selection into treatment. In this section we discuss the possibility that some unobserved confounders exist and the methods we used to face this eventuality. As we said in chapter 2, when we could not assume that unconfoundedness hold conditional only to observed covariates, but this assumption would be tenable only conditioning also on some unobservables, we term the assignment mechanism as latent regular. In other words, selection is present also on unobservables.

In the context of our application, childbearing is in this case endogenous despite controlling for observed characteristics. There are several reasons why childbearing might be endogenous with respect to economic wellbeing. One obvious reason is that it is determined by adults' labour earnings. Given unobserved ability levels, fertility decisions are endogenous with respect to women's work decisions and therefore their earnings. Ability is an unobserved confounder inasmuch as it influences also take-up of modern contraception (Kim and Aassve, 2006). Also fecundity can represent an omitted variable in our application. As noted by Kim et al (2005), more fecund women are at risk of low investment in human capital. As a consequence, they earn less. Obviously, on the other hand, fecundity sensibly influences fertility.

The instrumental variable approach has been the workhorse in economics to handle the omitted variable problem, but relies of course on the fact that valid and relevant instruments are available. Whereas these are not always easy to come by we propose two alternatives in our setting.

The first is a variable that takes value 1 if the household has no male children in the first wave - 0 otherwise. This kind of instrument is widely used (e.g., Angrist and Evans, 1998; Chun and Oh, 2002; Gupta and Dubey, 2003). The argument is that couples have certain gender preferences for their children - in particular they tend to have a preference for having at least one son. In other words, couples are more likely to have another child if the previous ones were

girls. In so far couples have a preference for boys, such a variable work well as an instrument since it is expected to have an impact on fertility but not a direct effect on poverty. Hence, the exclusion restriction seems to be reasonable. The strong preference for sons in Vietnam is confirmed by many studies (Haughton and Haughton, 1995; Johansson, 1996 and 1998; Belanger, 2002). Also monotonicity seems to be plausible with this instrument, since the presence of defiers is unlikely. In fact, the no-defiers assumption implies that households that would have (at least) one child between the two waves if they had one or more male children in the first wave (that is, no "encouraged" to have more children, $Z = 0$) would also have more children if they had no male children ("encouraged" to have more children, $Z = 1$). Moreover, the instrument can be thought of as being randomised, since households can clearly not choose the sex of their children[31]. To better highlight the preference for sons we selected only households that had at least 2 children in the first wave. Thus, having only girls in the first wave, can proxy an exogenous source for increased fertility between the waves.

The second instrument is a variable equal to 1 if in the community where the couple reside no contraceptive method, between IUD and condom, is available and 0 otherwise[32]. Instruments based on geographical variation in availability of services are not new (see for example: McClellan et al., 1994; Card, 1995). The variable we propose works well as an instrument if households living in communities with no contraceptive facilities have higher risks of childbearing and if contraceptive availability in the community has no direct effect on consumption growth. However, it is not unlikely that community characteristics that impact on the availability of contraceptive can also have an effect on households' poverty. Therefore, we cannot necessarily assume this instrument to be completely randomised as we do for the first one. Randomisation can only be assumed in so far we are also conditioning on a set of

[31] This is not completely true in those countries were the selective abortion are a current practice. It was found that this is the case for example for India where amniocentesis diagnoses are available and used for sex-selective abortions (Gupta and Dubey, 2003).
[32] IUD and condom are the most available contraceptives method in Vietnam and the IUD is the most largely used (Anh and Thang, 2002).

background variables. Controlling for covariates is consequently important in this setting. Monotonicity seems plausible also in this case, as the presence of defiers seems unlikely. In this case, it would imply that households who would have (at least) one child between the waves if they live in a community with available contraception (that is, not "encouraged" to have more children, $Z_i = 0$) would also have one child if they live in a community without contraception ("encouraged" to have more children, $Z_i = 1$).

5. 5 Instrumental variable methods results

The fact that the first instrument is randomized means that we can apply the Wald estimator without covariates. The results, presented in table 5.13, indicate a strong and negative effect of new children on the consumption expenditure and comparing it with the Frolich estimate we can see that the results are similar, confirming that controlling for covariates does not make a huge difference[33]. This result can be seen as reinforcement of the fact that the instrument can be assumed as randomised.

Two important considerations are in order here. First, since we selected households with at least 2 children these estimates refer properly only to this sub-population. Moreover, and more importantly, since IV estimates the LATE, these results are referred properly only to the latent sub-population of compliers, who are those choosing to have another child, *because* they did not already have a male child. The extent to which estimates can be compared with PSM estimates depends, in general, on the nature of the instrument. If we can assume that the average causal effect for "currently" non-compliers (always-takers and never-takers) is equal to the average causal effect for "currently" compliers (the LATE) then LATE and ATE coincide. This hypothesis is a-priori strong[34].

[33] The Frolich estimates have been obtained as the ratio of two matching estimators. We used kernel based matching. The final point estimates and standard errors were obtained bootstrapping over 1000 iterations. The covariate we use in the IV procedures are the same as those used in the analyses presented in section 5.2.

[34] This assumption is satisfied, for example, in case of homogeneous effects.

Table 5.13 – Local average treatment effect estimates through Instrumental Variables

Estimates	Instrumental variable			
	Son preference		Contraception availability in the community	
	Wald estimator	Frolich estimator	Wald estimator	Frolich estimator
LATE (standard errors in parenthesis)	-429 (228)	-490 (630)	-8822 (8597)	-785 (1672)
Proportions of compliers	0.28	0.16	0.09	0.01

Notes: point estimates and standard errors for the Frolich's method have been obtained by bootstrapping over 1000 iterations. The covariates used in the Frolich estimator are the same as those used in the analyses presented in section 5.2.

However, in the case where we use the sex ratio of the children as the instrument, one may argue that LATE and ATE will indeed be quite similar.

First, we note that the estimated proportion of compliers equals 0.2, which is a quite high compared to many other studies (see e.g. Angrist and Evans, 1998). This proportion of compliers is equal to the estimated causal effect of the instrument on the treatment variable D, $E[D_{1i} - D_{0i}]$. Hence, there is evidence to suggest this instrument being strong. Moreover, the fact that male preference is likely to be a nationwide phenomenon in Vietnam implies that the estimated effect on the sub-group of compliers could be referred to the whole population. This argument is of course supported by the fact that the LATE from the IV estimation (-429) is in this case almost equal to the ATE from the PSM estimation (-411)[35]. In this sense the IV can be viewed, in this special case, as a robustness check for the estimates resulting from the methods based on the UNC.

Using the community level availability of contraception gives a very different picture. The first issue is that we cannot assume that this instrument is truly randomised, which means that controlling for other covariates is essential. As a confirmation of this fact we observe a huge difference between the Wald

[35] We contrasted the LATE with the ATE instead than ATT because in the estimation of LATE we use either treated and untreated compliers, likewise in the calculation of ATE both treated and controls (in the whole population) are included.

and the Frolich estimates. Frolich estimates, which are more reliable, show a strong and negative effect for the sub-population of households that are "encouraged" to have a child by the lack of contraception in the community. This LATE is not comparable with the LATE estimated with the first instrument (i.e. using the sex ratio) since the sub-population of compliers are very different.

The estimated proportion of compliers for the second instrument is 0.01, hence the sub-population of households that "reacts" to this instrument is small. One could argue that since in Vietnam contraception is quite diffused compliers are in this case likely to be a rather selected group of marginalised households. This also could contribute to explain the much stronger effect. The low proportion of compliers implies that this instrument is weak and hence care is needed in the interpretation of this result since, as noted by AIR, the sensitivity of IV estimator to violations of exclusion restriction and monotonicity is higher when the proportion of compliers is lower. We have to note that IV estimates are quite imprecise with high standard errors, especially for this second instrument.

Whereas these considerations would favour the sex ratio as an instrument over and above the community level availability of contraception, it is important to bear in mind that the latter has a clear policy relevance. In fact, contraception availability in the communities is a variable on which policy makers could act. The Frolich estimates indicate that that the expected causal effect from a fertility reduction induced by raising contraception availability in the communities is quite high. However, the size of the sub-population reacting to this policy (compliers) is rather small. Obviously, policy makers should consider both aspects in order to calibrate efficient policies. In our case, the policy could have a huge effect on a very small group.

The other estimates also indicate that fertility impact considerably on economic wellbeing, but the policy implications are less direct. If policy makers worry about large households with many children, then targeted tax reductions and other benefits could help. In this way the existing gap between households with different number of children will be reduced. This policy clearly does not act on poverty through a fertility reduction and instead could influence a raise in fertility. However, our estimates do not generally suggest that the level of

economic wellbeing will increase fertility levels (see Table 5.3). In particular, the expenditure levels as recorded in the first wave do not seem to have any strong income effect on childbearing decisions[36].

5.6 Concluding remarks

Several approaches are available in order to estimate causal effects. The appropriateness and interpretations of these models depend on the application at hand, and importantly, the available instruments. In many cases, methods relying on the UNC assumption are chosen, simply because instruments are hard to come by. The various implications of these methodological choices are rarely considered in applied work, but, as we point out, the underlying assumptions are important, especially when there is interest in comparing estimates from different methods. We discuss these methods in light of an application where we consider the effect of fertility on changes in consumption expenditure. The issue is that childbearing events cannot be considered as an exogenous measure of fertility, especially when the outcome relates to economic wellbeing – in our case measured in terms of consumption expenditure. However, the discussion of the methods is general and applies to many other applications.

Using methods based on the UNC assumption, such as simple linear regression and propensity score matching, we find that those households having children between the recorded waves have considerably worse outcomes in terms of changes in consumption expenditure. The negative impact is, however, highly heterogeneous, and varies substantially with education, for instance. We then assess, through an extensively sensitivity analysis, the potential effect from omitting relevant, but unobserved, variables without actually implementing an Instrumental Variable approach. This is a very useful tool especially when instruments are not available. In our application the estimates are robust with

[36] However, the effect of consumption in 1992 on the childbearing decision reported in table 5.3 refers to the whole sub-sample and we cannot exclude that some heterogeneity effect is in order. For example, for some quintile of the consumption distribution the effect could be nonzero.

respect to the omission of unobserved omitted variables. We find that the estimated effect becomes non-significant only if the association between the omitted covariate, selection and the outcome is extremely (and unreasonable) large.

Despite the robustness of the UNC, in our application we implement nevertheless the IV method using two different instruments. The first is a well-used instrument that relates to couples' preference for sons. In this case the instrument is a binary variable taking value 1 in those households that at the first wave had no male children and 0 otherwise. The IV estimation is implemented for sub-sample of households with at least two children in the first wave. Since the instrument is close to being randomised, a simple Wald estimator can be used. The second instrument takes value 1 for households residing in a community where none of the contraceptive methods IUD and condom was available at the first wave and 0 otherwise (at least one was available). This instrument is not randomized and, hence, requires controlling for covariates. Whereas both instruments are reasonable they provide, not unexpectedly, very different parameter estimates. The fact that the IV estimates the LATE, as opposed to the ATE and ATT, is the key reasons for these differences.

The use of Instrumental Variable methods in our application illustrates that reasonable instruments can lead to estimates that differ from those of methods based on UNC but also differ among them. In fact, compliers for one instrument can be very different from compliers to another instrument and, consequently, if the treatment effect is heterogeneous the estimated LATE in the two cases will necessarily differ. With the first instrument we estimated a negative impact of fertility on poverty with a magnitude not dramatically different from that obtained by method based on the UNC. This could be an effect of the fact that the preference for son is quite a general phenomenon in Vietnam not involving particular kinds of households. The estimated proportion of compliers in this case is actually quite high: 20%. The estimate with the second instrument, on the contrary, is much higher, in absolute value. The estimated proportion of compliers in this case is small: 1%. This small sub-population of households reacting to the availability of contraceptives is likely to

be highly selected. These households live in areas where no contraceptives were available. Clearly their opportunity to control fertility through contraceptive practices is much reduced, as it is unlikely that compliers are able to get contraceptives from elsewhere. In this sense, these households have a higher exposure to undesired childbearing. These communities are also likely to be more disadvantaged compared to others.

Whereas the estimates based on this instrument is very different compared to the one based on the sex preference, an advantage is that it does have direct policy relevance, simply because the instrument itself is a policy variable. The effect on this sub-population is high and importantly, much higher than what is estimated for the whole population through the ATT and ATE. However, the size of this sub-population is rather small, which is an equally important consideration for the policy maker.

Appendix to chapter 5

Table A.1 – Description of the covariates used in the applications presented in this chapter (all measured in 1992)

Label	Description
Household level covariates	
Sexhhh	Sex of the household head: Male = 1; Female = 0
Kinh	Kinh (principal ethnic group)= 1; non-Kinh = 0
perkids_04	Percentage of kids aged [0-4] out of total number household members
perkids_59	Percentage of kids aged [5-9] out of total number household members
perkids_1014	Percentage of kids aged [10-14] out of total number household members
permale_1545	Percentage of male aged [15-45] out of total number household members
perfema~1545	Percentage of female aged [15-45] out of total number household members
Farm	Farmer households = 1; Non-farmers = 0
Edu	Educational index obtained as: $\sum GC_i / \sum GP_i$ where GC_i = n. of completed grades by the household member i GP_i = n. of potential grades the member i could complete at a given age. This is given by: $GP_i = \begin{cases} 0 & \text{if } age(i) < 7 \\ age(i) - 6 & \text{if } 7 \le age(i) \le 18 \text{ age}(i) \\ 12 & \text{if } age(i) \ge 18 \end{cases}$.

Educational grades we considered are:
Primary school (compulsory): grades 1-5
Intermediate: grades 6-9
High school: grades 10-12.
Note that each grade corresponds to one year and primary school starts when children are 6 years old.

rlpcex1 Equivalised household consumption expenditure

Agehhh Age of the household head

Hhsize Household size (total number of members)

peractm_1545 Percentage of working male aged [15-45] out of the total number of males aged [15-45] in the household

peractf_1545 Percentage of working female aged [15-45] out of the total number of females aged [15-45] in the household

Community level covariates

IEI Infrastructural Economic Index obtained as:

IEI= road + elect + pipeb + cradio + post + marketd + marketpdc + pubtrpass + bigent + usefert+ agrext + tottrac

where:

road = 1 if a road is present; 0 otherwise
elect = 1 if electricity is available; 0 otherwise
pipeb = 1 if pipe born water is available; 0 otherwise
radio = 1 if a radio station is present; 0 otherwise
post = 1 if a post office is present; 0 otherwise
marketd = 1 if a daily market is present; 0 otherwise
marketp = 1 if a periodic market is present; 0 otherwise
pubtras = 1 if public transport is available; 0 otherwise
bigent = 1 if a big enterprise is present; 0 otherwise
usefert = 1 if majority of famers use chemical fertiliser; 0 otherwise
agrext = 1 if an agriculture extension centre is present; 0 otherwise
tractors = 1 if at least one large tractor is available; 0 otherwise

EDI	Educational Index obtained as: EDI= prims + lowsecs + upsecs + clitprog. Where: prims = 1 if a primary school is present; 0 otherwise lowsecs = 1 if a lower secondary school is present; 0 otherwise upsec = 1 if a upper secondary school is present; 0 otherwise clitprog = 1 if a community literacy program was implemented in the 5 years period before 1992; 0 otherwise
HFI	Health Facilities Index obtained as : HFI = hosp + clinic + comcli + doct + phar + phydoct+ nurse + pharma + tmidw. Where: hosp = 1 if an hospital is present; 0 otherwise clinic = 1 if a private clinic is present; 0 otherwise comcli = 1 if a communal clinic is present; 0 otherwise doct = 1 if at least one doctor is present; 0 otherwise phar = 1 if a pharmacist is present; 0 otherwise nurse = 1 if a nurse is present; 0 otherwise pharma = 1 if a pharmacy is present; 0 otherwise tmidw = 1 if a trained midwife is present; 0 otherwise
Regional dummies	
region1	Northern Mountains
region2	Red River delta
region3	North Central Coast
region4	South Central coast
region5	Central Highlands
region6	Southeast
region7 (reference category)	Mekong river delta

Chapter 6
The multilevel dimension in the estimation of the causal effect of fertility on poverty in Vietnam

Introduction

In this chapter we re-analyse the estimation of the causal effect of fertility on poverty, treated in the previous chapter, with the goal of keeping into account the multilevel dimension of the problem. The methodological framework to which we refer was developed in chapter 3, where we discussed the statistical aspects of causal inference in a multilevel setting. In chapter 1, we introduced the substantive motivations that push us to emphasize the multilevel data structure in our application. In the present chapter we report the results of our analyses. In section 6.1 we analyse the effect of fertility on poverty using multilevel models. In section 6.2 we adopt a different strategy consisting in a combination of multilevel models for the estimation of the propensity score and matching methods. In section 6.3 we address the complication due to the potential violation of the SUTVA in a multilevel setting.

6.1 Estimating the effect of fertility on poverty using multilevel models

In order to study the effect of fertility on poverty keeping into account the clusterisation of households in communities we can use a two-level model. The outcome and the treatment of interest are defined as in the applications presented in the previous chapter; the same sub-sample selection criteria have been applied. Therefore, also in this chapter we use the sub-sample of households with at least one married female aged [15-40] in the first wave.

We started by estimating a null model, that is, a model without covariates, in order to verify if there is a sufficient between-clusters variability to justify the use of a multilevel model. As we can see from table 6.1, the second level variability is highly significant in the null model. This is suggested by the Likelihood Ratio (LR) test on the random intercept of which we report the p-value[37]. The other models in table 6.1 include covariates at both levels of the structure. We are mainly interested in the effect of a specific first level dummy covariate: *D*, representing, as in chapter 5, the occurrence of at least one childbearing event between the two waves in the household; the other covariates play the role of control variables. These covariates are the same as those used in the regressions and matching procedures presented in the chapter 5.

In table 6.1 we resume results of four models. As far as the random part is concerned, models 1 and 2 include only a random intercept, while models 3 and 4 specify also a random slope for *D*. This is interesting since it allows us to study the community heterogeneity in the treatment effect which is one of the point of interest in causal inference in a multilevel setting, as said in section 3.2. From the LR tests we can see that the variance of the random intercepts are always highly significant, while those of random slopes in models 3 and 4 are not.

[37] We halved the p-values as suggested by Snijders and Bosker (1999) to keep into account the problem we discussed in section 3.3 (pag. 93).

Table 6.1 – Multilevel linear models for the estimation of the fertility effect on poverty

The characteristics of the models

Models	D	$\overline{D}$	Covariates	RI	RS
Null	no	no	no	yes	no
1	yes	no	yes	yes	no
2	yes	yes	yes	yes	no
3	yes	no	yes	yes	yes
4	yes	yes	yes	yes	yes

The estimates

Models	Fixed effects		Random effects (Standard deviations)		LR tests (p-values)	BIC
	D	$\overline{D}$	RI	RS		
Null	no	no	489 (45)	no	RI: 0.0000	25925
1	-388 (73)	no	337 (33)	no	RI: 0.0000	24287
2	-374 (74)	-366 (302)	363 (41)	no	RI: 0.0000	24259
3	-385 (70)	no	303 (34)	135 (101)	RI: 0.0000 RS:0.5231	24295
4	-377 (68)	-375 (304)	358 (35)	141 (97)	RI: 0.0000 RS:0.6000	24288

Notes: RI = random intercept; RS = random slope; BIC = Bayesian information criterion. Standard errors are in parentheses.

Models 2 and 4 differ from models 1 and 3 because they include an additional covariate, $\overline{D}$, which represents the cluster mean of *D*. As mentioned in section 3.3, in this way we can distinguish the within and the between effects of the treatment of interest.

Form table 6.1, we see that in this application the between and within effects are not significantly different. In fact, the estimates of the coefficient of *D* obtained using models 1 and 3, that mixes up the between and within effects, are very similar to those of models 2 and 4, which are pure between effects. We can see this more directly from the coefficient of $\overline{D}$ in models 2 and 4, which represent the difference in the between and within effects. This coefficient is not significantly different from zero.

Globally, from the estimates presented in table 6.1 we see that the negative, strong and significant effect of childbearing events on poverty that we found in the previous chapter is confirmed. As expected, the estimated effect is in line to the one obtained from the single level regressions. As in the analysis presented in chapter 4, the between-community variability in the consumption growth pattern is highly significant even after controlling for first and second level covariates. However, we found no evidence of a significant between-community variability in the treatment effect as well as of a difference in the within and between effects.

As noted in chapter 3, when we are interested in the study of causal effects in a multilevel setting, multilevel models can be very useful but, because of some pitfalls we have explained, we prefer to combine or substitute them with matching methods. One of these drawbacks is that multilevel models impose a functional form relating the outcome and the treatment. Moreover, in this section we have not posed two important issues we analyse in the following sections: the multilevel nature of the selection process and the potential violation of the SUTVA.

6.2 Propensity score matching based causal inference: comparing different strategies

In this section we explore the propensity score matching method for the estimation of causal effects in a multilevel setting. As extensively said in section 3.5, when we use a PSM approach in a multilevel framework we have to explicitly consider the multilevel nature of the selection process. In this section we compare two strategies. The first consist to specify a multilevel model for the propensity score and to implement the matching on the estimated empirical bayes probabilities resulting from this model. The second consists in two stages: first we estimate a multilevel model for the probability of receiving the treatment and obtain the empirical bayes predictions of the random effects included in this

model; then, we include these predictions as additional covariates in the estimation of a single level propensity score model.

In the table 6.2 we describe the propensity score models we compare in the following. Models 1 and 2 are simple single level logit models. The difference between them is that the first model includes both household and community level covariates, while the second includes only covariates at the household level. The reason for considering also models without community covariates will be explained in the sequel.

The second group of models (3-8) collects two-level logit models. They differ for the inclusion or exclusion of community covariates and for the specification of the random part; some of them include one random slope for very important covariates. The last group of models (9-12) includes single level logit models specified in accord to the two-stage procedure outlined before and discussed in chapter 3. They are basically single level logit models including some particular additional covariates: the empirical bayes predictions of random effects obtained after the estimation of the two-level logit models indicated in parentheses. We are not primarily interested in the goodness of fit (reported in the table 6.3)[38] of these models, but in the balancing they allow us to achieve. Likewise in chapter 5, as a measure of the balancing of the covariates we adopt the absolute standardised bias (ASB). As we can see from tables 6.3 and 6.4, a better fit, represented here by a lower BIC, not always correspond to a better balance, represented here by a lower ASB. We want to stress three types of comparisons among these models:

1) Comparison among single level and two-level propensity score models;
2) Comparison among the second and the third group of models;
3) Comparison among models ignoring community level information.

[38] In fact, the main purpose of the propensity score is not to predict participation in the treatment but balance all observed covariates in the matching procedure (Augurzky and Schmidt, 2001). Therefore, we are not interested in the goodness of fit of model specification but in balancing, that we assess through the ASB. Moreover, "perfect" prediction should be avoided, since if $P(D=1|X)=1$ or $P(D=1|X)=0$ for some value of X, we cannot match on these values of X as they are out of the common support.

Table 6.2 – Description of the propensity score specifications we compare

PS	Description
I - Single level logit models	
1	With X and C
2	With X, without C
II - Two-level logit models	
3	With X and C; RI
4	With X, without C; RI
5	With X and C; RI; RS for *kinh* (principal ethnic group)
6	With X and C; RI; RS for *farm* (binary indicator for farmer households)
7	With X and C; RI; RS for *edu* (index for educational level)
8	With X, without C; RI; RS for *edu*
III - Single level logit models with predicted random effects included as additional covariates	
9	With X,C and predictions of RI (obtained from model 3)
10	With X,C and predictions of RI and RS for *edu* (obtained from model 7)
11	With X, without C and with predictions of RI (obtained from model 4)
12	With X, without C and with predictions of RI and RS for *edu* (obtained from model 8)

Notes: PS = propensity score specification; X = household level covariates; C = community level covariates; RI = random intercept; RS = random slope

The first comparison is among single-level versus two-level logit specification for the propensity score model. The expected benefit from the second group of models (3-8) with respect to the first one (1-2) is that a multilevel specification allow us to keep into account the unobserved community characteristics.

In the previous chapter we have seen that the single level specification we have adopted for the estimation of the propensity score ensured us a satisfactory balance in the first as well as second level covariates. This is confirmed in the table 6.4 where we see that the mean and median ASB for the first and second level covariates included in model 1 are quite low. Let compare model 1 versus models 3, 5, 6 and 7. As we can see from the table 6.4, multilevel propensity score models show worse balance in the first level covariates (X) with respect to the single level model 1. On the contrary, the balancing for the second level covariates is often better with the two-level models.

Table 6.3 – Some propensity score specifications: significance of random effects and goodness of fit (standard errors are in parentheses).

Models	Random effects (Standard deviations)		LR tests (p-values)	BIC
	Random Intercept	Random Slope		
1	no	no		1851
2	no	no		1810
3	0.290 (0.120)	no	RI: 0.0040	1830
4	0.320 (0.110)	no	RI: 0.0036	1813
5	0.290 (0.120)	0.001 (0.360)	RI: 0.0037 RS: 0.0004	1837
6	0.210 (0.200)	0.300 (0.240)	RI: 0.0030 RS: 0.0004	1837
7	0.160 (0.320)	0.004 (0.160)	RI: 0.0028 RS: 0.0005	1839
8	0.160 (0.360)	0.050 (0.003)	RI: 0.0015 RS: 0.0600	1821

Table 6.4 – Comparison among different propensity score specifications: balancing and estimates.

Propensity score	Absolute Standardised Bias after matching						Estimates	
	First level covariates		Second level covariates		Predicted random effects			
	Mean	Median	Mean	Median	RI	RS	ATE	ATT
I - Single level logit models								
1	3.7	2.9	4.5	4.7	no	no	-411	-356
2	5.4	4.4	10.4	9.2	no	no	-421	-351
II - Two-level logit models								
3	6.0	4.8	3.6	2.8	6.9	no	-492	-431
4	7.0	6.2	3.4	3.2	6.4	no	-541	-470
5	5.5	3.5	1.5	1.2	6.8	5.9	-434	-384
6	5.7	3.9	4.2	3.8	8.5	7.6	-464	-514
7	5.7	5.7	4.4	4.7	9.8	8.2	-450	-375
8	6.3	5.3	6.8	3.8	7.5	8.7	-447	-407
III - Single level logit models with predicted random effects included as additional covariates								
9	3.7	3.1	3.9	4.1	4.1	no	-476	-401
10	3.1	2.9	3.7	4.0	3.6	4.0	-482	-429
11	3.6	2.7	3.7	3.4	2.0	no	-467	-418
12	4.2	3.2	3.5	3.1	3.7	3.9	-454	-423

This comparison shows that using a multilevel model for the propensity score could represent a danger if we do not take carefully into account what happens to the balancing of observed covariates. We tried also other specifications, including some interactions, higher order terms, first level centred covariates, and so on. The results obtained are similar to the ones showed here.

At this point we consider our proposed two-stage procedure (the third group of models in the table 6.2). As we can see from table 6.4, this strategy allow us to maintain a good balancing, on average, in both first and second level covariates. Moreover, in this way we balance also the predicted random effects well, while with models of the second group this is not the case. Since they capture unobserved community characteristics we can think that this strategy is the one to be preferred, because it allows to balance observed covariates and potentially important unobserved community variables. This seems to be confirmed by the last two models (11 and 12). These models, as well as model 2, 4 and 8 ignore the community level information. The reason to consider these models is that we want to see what would be the balancing of the community characteristics using the different propensity score specifications, in case no community level variables were observed. We get an interesting result. Models 11 and 12 allow to achieve a reasonably good balancing of community variables, even if these are not included in the matching set. Also results from models 4 and 8 are good. On the contrary, a single level propensity score model that includes only first level covariates (model 2) does not guaranty an acceptable balance of second level covariates.

Summarizing, we can say that the inclusion of predicted random effects in the single model for the propensity score, in our application, actually "substitute", in a way, the inclusion of second level covariates. A similar reasoning hold for multilevel specifications of the propensity score with respect to a single level one. This result is potentially very important for those situations, not uncommon in social studies, where data available contains no, or insufficient, information at the cluster level. Concluding, model 10 seems to "dominate" all the others and is chosen as the "best" specification for the propensity score.

However, these results cannot be directly applied to other situations. As we already said in chapter 3, we reserve a formal investigation on the performances of the different strategies proposed to future work. It must be noted that, in general, the specification of the propensity score is an "art" that require a strong degree of adjustment to the situation at hand. For example, we noted that in order to achieve a good balancing in all observed covariates it is often needed to include some interactions and higher order terms. In the model specification we are essentially driven by the balancing we are able to obtain. In other words, we do not know a priori which interactions and higher order terms are to be included. In a similar way, it can be argued that, *a priori*, we cannot know if a single level or a multilevel specification for the propensity score has to be preferred in a given application. However, we think that a simulation study could reveal under which conditions a multilevel specification is expected to give better balancing.

As the estimates are concerned, we can see from table 6.4, that ATT and the ATE obtained through a PSM procedure are not noticeably sensitive to the specification of the propensity score. Estimates based on the propensity score 10, which we choose as a reference, substantially confirm our previous results. The estimates, nevertheless, are a bit higher, in absolute value, than those obtained using a single-level propensity score model. This result is qualitatively in line with those obtained by the other models of the third and second group, indicating that controlling for potentially unobserved community level confounders, we get a slightly stronger estimated effect.

In this respect, it is interesting to reconsider the sensitivity analysis to potential unobserved confounders we presented in section 5.3.4. This analysis has showed that our estimates are quite robust, but for several combinations of the association between the unobserved simulated confounder, the treatment indicator and the outcome, the estimated ATT is stronger than the baseline. Results obtained in the present section seem to agree with this trend, meaning that if some unobserved community confounders was in action this should likely imply a stronger estimated causal effect, leaving unchanged, qualitatively, previous results. Our proposed two-stage procedure, as we anticipated in chapter

3, can be seen as a variant of the sensitivity analysis we implemented in chapter 5. Here, we are not simulating an unobserved confounder, but we are predicting random effects, which are expected to represent unobserved cluster level characteristics. Likewise we did in the sensitivity analysis for the simulated confounder, also the prediction of the random effects are included in the propensity score as additional covariates. We have seen that the estimates we get, by matching on the resulting propensity score, are not dramatically different from the reference (that is, those obtained using a single level propensity score). Therefore, we can conclude that the PSM procedure presented in section 5.2 seems to be not sensitive to departures from UNC due to unobserved cluster level confounders. This is in line with the results of the sensitivity analysis carried on in section 5.3.4.

6.3 Estimation results under a weaker version of the SUTVA

In this section we compare the estimates obtained under the standard version of the SUTVA used in all the previous analyses with those we obtain under the weaker version outlined in the section 3.5. We briefly remember that this weaker version of SUTVA assumes no interference among households living in different communities. On the other hand, to keep into account potential interference among households belonging to the same community we introduce the binary indicator, L, taking value 1 for households living in communities with a "high" level of childbearing events (treated) and 0 otherwise. To go ahead in the analysis, we need to empirically distinguish between "high" and "low" level of treated community. We used the following criterion. We calculated the proportion of treated households in each community. Then, we assigned the value 1 (high) to communities whose estimated proportion is significantly (at 5% level) higher than the national average.

As declared in section 3.5, as soon as we weaken the SUTVA it is natural to consider more causal estimands of potential interest then under the standard version of this assumption. They refer to the effects of two treatments: *D,*

operating at the household level, and *L,* operating at community level. We are mainly interested in the treatment *D* in this application. The causal estimands referred to *D* can be defined as follows in our context:

ATE^{D} : is the average causal effect of childbearing events calculated on the whole population,

$ATE^{D}_{|L=1}$: is the average causal effect of childbearing events calculated only for households living in community with a "high" level of treated,

$ATE^{D}_{|L=0}$: is the average causal effect of childbearing events calculated only for households living in community with a "low" level of treated.

Obviously, we can consider also the ATT versions of these parameters, that is, the corresponding parameters calculated only on the sub-group of treated households. We are interested in the comparison between the effect of childbearing in community with high versus low fertility, as well as in the comparison between the estimated causal effect under the standard and the weaker version of the SUTVA.

In order to estimate these parameters, we used a PSM method. We employed model 10 of table 6.2 as the specification of the propensity score. In order to calculate $ATE^{D}_{|L=1}$ and $ATE^{D}_{|L=0}$ we separately estimated the propensity score models, respectively, for households residing in communities with a "high" and a "low" level of treated. The matching method employed was always the nearest neighbour matching.

As we see from table 6.5, the estimated ATE and ATT for treatment *D* under the two versions of the SUTVA are quite similar. This confirms that, globally, the within and between effects are not significantly different, as already indicated from estimates of multilevel models presented in table 6.1.

Table 6.5 – Estimated causal effects of childbearing events under two versions of the SUTVA

ATE		ATT	
Estimates obtained under the standard version of the SUTVA			
ATE^{D}	-482 (92)	ATT^{D}	-429 (99)
Estimates obtained under the weaker version of the SUTVA			
$ATE^{D}_{\|L=1}$	-420 (146)	$ATT^{D}_{\|L=1}$	-566 (176)
$ATE^{D}_{\|L=0}$	-447 (113)	$ATT^{D}_{\|L=0}$	-313 (107)
ATE^{D}	-440 (90)	ATT^{D}	-425 (100)

Notes: The estimates are all obtained using a PSM procedure following the specification used in model 10 (see table 6.2). The matching method employed is the nearest neighbour.

The average causal effect of childbearing on poverty in communities with a "high" level of childbearing events is not radically different from the parameter calculated in the remaining communities. On the contrary, if we condition on households that had at least a child between the two waves the situation is different. In fact, the $ATT^{D}_{|L=1}$, the average causal effect of fertility on poverty for treated households living in high-fertility communities, is higher, in absolute value, than $ATT^{D}_{|L=0}$. This result can be due to the competition for and share of resources that is in action within communities. More explicitly, the facilities, services (e.g. provided by health care or family planning centres), benefits (e.g. maternity benefits), which are made available in a community for help households with children are subject to economic limits. In communities where the number of children is higher these constraints generate competition among households. It is likely that in communities with more childbearing events, some households cannot gain some benefits, or to obtain some services are forced to move to other communities or have to pay private providers.

This result is very important for policy making. Despite the fact that the effect of childbearing is negative for both types of communities, its impact is stronger in high-fertility communities. This suggests that policy interventions are more pressing in that type of communities. For example, policy maker can be

encouraged by these results to improve those facilities and increase benefits in communities where the number of childbearing events is high.

As far as the effect of the variable *L per se,* we estimated the ATE^L to be equal to -101 with a standard error of 162[39]. Therefore, it seems that living in a community with a "high" level of fertility, here proxied by the childbearing events occurred between the two waves, or in a community with a "low" level of fertility is not different for the households' living standard. This, obviously, after controlled for compositional and contextual differences existing among communities, captured by the control variables. Therefore, the negative association we often find between the fertility rate in the place of residence and living standards is likely to be due to the association among the level of fertility in a geographical area and other socio-economic characteristics (economic development, infrastructures, culture, and so on). In other words, we found that at the individual level (household level, in our case) fertility has a negative causal effect on wellbeing, while at the aggregate level (community) there is a mere (spurious) association.

6.4 Concluding remarks

In this section we qualitatively confirm the main result obtained in chapter 5. The effect of childbearing events on poverty is confirmed to be negative, strong and significant also keeping explicitly into account the multilevel dimension of the problem.

Using multilevel linear models we found, likewise in chapter 4, that the households' consumption growth shows a high degree of variation between communities confirming the importance of the multilevel dimension. On the contrary, there is no evidence for a significative between-communities variation in the childbearing effect. However, when we use a weaker version of the SUTVA allowing interference among households living in the same community,

[39] We calculated also the conditional versions of this parameter, as well as the ATT version. Results confirm the non-significance of the effect of the variable *L*.

we found that the average causal effect is stronger in high-fertility communities then in low-fertility ones. On the other hand, we found no significative effect of living in a high-fertility community versus a low-level one.

From a methodological point of view, we found that our suggested strategy for the specification of the propensity score model, consisting in the inclusion of predicted random effects as additional covariates in a single level logit model, is the method to be preferred at least in this application. In fact, this strategy allowed to achieve a good balance in both first and second level observed covariates, as well as, in the predicted random effect capturing potentially important unobserved community level confounders.

An important result, useful in those situations where no cluster-level information is available, is that our strategy allowed to balance community-level variables even when these were not included in the matching set for the estimation of the propensity score model. This indicates that the predicted random effects "substitute" the cluster-level variables and, hence, achieving a good balance for the former ensures a good balance also for the latter. However, in future work we are planning to use simulation analysis to understand the conditions under which the results we found in our application can be generalised.

Final remarks

In this work we re-analyzed the relationship between fertility and poverty, which is a long contested issue among demographers and economists. We tried to contribute to this literature by using a proper causal inference approach and adequate panel data concerning Vietnam.

We started the work by discussing the determinants of poverty and fertility, as they emerge in the theoretical and empirical literature on the topic. This discussion was very useful in the stage of the choice of the control variables to include in our analyses. From the literature, as well as from our analyses, the role of education emerges, prominently. We found, in accord to the literature, that households with educated members show, *ceteris paribus*, a higher fertility and a lower consumption expenditure.

In chapter 1, we motivated the two key perspectives we used in this research: causal and multilevel. A causal approach is needed in order to give sound policy advice. A multilevel approach is, on the other hand, adequate since we see as crucial to explicitly take into account that the characteristics of the place where households live sensibly influence both studied phenomena. Therefore, there is a two-level data structure – households clustered in communities – which has important statistical and substantive implications.

We argued that the data we use, coming from the Vietnamese Living Standard Measurement Survey (VLSMS), are sufficiently reach to allow us to take into account both perspectives. In fact, the VLSMS, which is a panel consisting of two waves, includes rich information on variables that are important determinants for the households' standard of living and fertility behaviour. For example, it collects data on education, employment, fertility and marital histories, together with detailed information on household income and

consumption expenditure. Moreover, a very interesting feature of the VLSMS is that it also provides, for the rural area, detailed community information from a separate community questionnaire. This allowed to include into the analyses important contextual information.

The longitudinal dimension of the data available was crucially important to allow us to draw robust causal inference about the effect of interest. In fact, as we discussed, by using data on two time points we properly implemented a pre-post treatment study which was vital for our study of causal inference.

In chapter 2, we presented the potential outcomes framework, which is the approach to causal inference we adopt in this work. We reviewed several methods for causal inference, stressing their differences with respect to the underlying assumptions and data requirement. In particular, we contrasted methods relying on the *Unconfoundedness Assumption* (UNC), such as regressions and propensity score matching, with methods allowing for *selection on unobservables*, such as the Instrumental Variable (IV) estimators. We stressed the fact that these methods are not equivalent in what they estimate. With Regressions and Propensity Score Matching (PSM) we can identify and estimate the Average Treatment Effect (ATE) and the Average Treatment effect on the Treated (ATT), while IV methods give the Local Average Treatment Effect (LATE), unless we are willing to impose very stringent additional assumptions. Since LATE is the average causal effect of the treatment on the sub-group of compliers, it is generally different from ATE and ATT. Moreover, different instruments identify the effect on different groups of compliers giving different estimates of LATE. A problem for policy making is that the compliers are in general an unobserved sub-group. However, we argued that IV methods estimate relevant policy parameter if the instrument itself is a potential policy variable. This review includes, in particular, recent methodology for using instrumental variables with covariates avoiding traditional methods, which often rely on strong assumptions. We assessed the issues outlined in this chapter with an application concerning the estimation of the causal effect of childbearing events on consumption expenditures growth, presented in chapter 5.

In chapter 3, we presented the general motivations for a multilevel research and the basic features of a multilevel linear model. We reviewed, in particular, the second level endogeneity problem. Then, we originally re-analysed the traditional multilevel models in the light of the potential outcomes approach, highlighting its pitfalls in the study of causal effects. Importantly, we discussed the main issues motivating the need to keep explicitly into account the multilevel dimension in a study of causal inference carried in a multilevel setting. The literature on this topic is very limited.

Our first contribution concerned the unitary discussion of some topics, which are usually treated separately. Moreover, we proposed a two-stage strategy for the specification of the propensity score in order to achieve a good balancing in the observed covariates, defined at each level of the hierarchical structure, as well as the predictions of random effects, which are entitled to capture unobserved higher level effects. At the first stage, we estimate a multilevel model for the probability of receiving the treatment. This includes a random intercept, as well as some random slopes capturing the between-cluster heterogeneity in the influence that some covariates have on the probability of being treated. Then, we calculate the empirical bayes predictions of these random effects. In the second stage, these predictions are finally included, as additional covariates, in a single level model for the propensity score. Finally, we discussed the potential violation of the SUTVA in a multilevel setting and a weaker version of this assumption, which addresses the interference existing among units belonging to the same cluster.

As the empirical part of the work is concerned, this is included in the last three chapters. Chapter 4 included an application of multilevel analysis techniques, which is interesting *per se.* In this analysis, a relevant residual between-community variability in poverty exit dynamics was found. This fact justified the use of multilevel techniques. Moreover, it allowed us to implement a study of communities' effectiveness, which is a topic common in educational research, but absent in poverty analyses. We suggested several use of the empirical bayes predictions of the random intercept, included in the model, for policy making. For example, we suggested a ranking, and a subsequent

categorisation of communities, in a "good" and a "bad" group, which can be used, along with a more qualitative analysis, like an intensive case study, to help policy maker to understand the contextual determinants of poverty exit and to individualize communities requiring more urgent intervention.

In chapter 5, we explored the use of several methods for the estimation of the causal effect of fertility on poverty. Using methods based on the UNC assumption, such as multiple linear regression and propensity score matching, we found that those households having children between the recorded waves have considerably worse outcomes in terms of changes in consumption expenditures. The negative impact is, however, highly heterogeneous and varies substantially with education, for instance. We then implemented, through an extensive sensitivity analysis, an assessment of the potential effects from omitting relevant but unobserved variables without actually implementing an Instrumental Variable approach. This is a very useful tool, in the sense that valid and relevant instruments are often hard to come by. In our application the estimates are robust with respect to unobserved omitted variables. We found that the estimated effect becomes non-significant only if the association between the omitted covariate, selection and the outcome is extremely (and unreasonably) large.

Despite the robustness of the UNC in our application, we implemented nevertheless the IV method using two different instruments. The first is a well-used instrument that relates to couples' preference for sons. Since the instrument is close to being randomised, a simple Wald estimator can be used. The second instrument relates to the availability of contraceptives in the community where the household resides. This instrument is not randomized and hence requires controlling for covariates. We found that the two instruments provide very different parameter estimates. The use of Instrumental Variable methods in our application illustrates that reasonable instruments can lead to estimates that differ from those of methods based on UNC but also differ among them. In fact, compliers for one instrument can be very different from compliers to another instrument and consequently if the treatment effect is heterogeneous the

estimated LATE in the two cases are likely to be dissimilar, as well as they are expected to be different with respect to the ATE and ATT.

With the first instrument we estimated a negative impact of fertility on poverty with a magnitude not dramatically different from that obtained by method based on the UNC. We argued that this could be due to the fact that the preference for son is quite a general phenomenon in Vietnam, not involving particular kinds of households. The estimated proportion of compliers in this case is actually quite high: 20%. The estimate with the second instrument, on the contrary, is much higher, in absolute value. The estimated proportion of compliers in this case is small: 1%. This small sub-population of households reacting to the availability of contraceptives is likely to be highly selected. Clearly their opportunity to control fertility through contraceptive practices is much reduced, as they are not able to get contraceptives outside the community family planning centres.

Whereas the estimates based on this instrument is very different compared to the one based on the sex preference, an advantage is that it does have direct policy relevance, simply because the instrument itself is a policy variable. The effect on this sub-population is high and importantly, much higher than what is estimated for the whole population through the ATE. However, we noted that the size of this sub-population is rather small, which is an equally important consideration for the policy maker.

In chapter 6, we re-analysed the causal effect of fertility on poverty, treated in the previous chapter, with the goal of keeping into account the multilevel dimension of the problem. Altogether, we qualitatively confirmed the main result obtained in chapter 5. The effect of childbearing events on poverty is found to be negative, strong and significant.

First, we analysed the effect of fertility on poverty using multilevel models. We found, likewise in chapter 4 for poverty exit dynamics, that the households' consumption growth shows a high degree of variability between communities, confirming the importance of the multilevel dimension. On the contrary, there is no evidence for a significative between-communities variation in the childbearing effect. However, when we use a weaker version of the

SUTVA, allowing interference among households living in the same community, we found that the average causal effect is stronger in high-fertility communities then in low-fertility ones. We argued that this result can be due to the competition for and share of resources that is in action within communities. More explicitly, facilities, services (e.g. provided by health care or family planning centres), benefits (e.g. maternity benefits), which are made available in a community for help households with children are subject to economic limits. In communities where the number of children is higher these constraints generate competition among households to get them. It is likely that in communities with more childbearing events some households cannot gain some benefits, or to obtain some services are forced to move to other communities or have to pay private providers.

This result is very important for policy making. Despite the fact that the effect of childbearing is negative for both type of communities, its impact is stronger in high-fertility ones. This suggests that policy interventions are more pressing in those areas. For example, policy makers can be encouraged by these results to improve key facilities and increase benefits in communities where the number of childbearing events is high.

On the other hand, we found no significative effect of living in a high-fertility community versus a low-level one, *per se*. This result confirms the idea that the negative association, we often find, between fertility rates in the place of residence and living standards is spurious, in the sense that it is likely to be due to the association among both phenomena at an aggregate level and other socio-economic characteristics (economic development, infrastructures, culture, and so on).

From a methodological point of view, we found that our suggested two-stage strategy for the specification of the propensity score model is the method to be preferred, at least in this application. In fact, this strategy allowed to achieve a good balance in both first and second level observed covariates, as well as, in the predicted random effect capturing potentially important unobserved community level confounders.

An important result, very useful for those situations where no cluster-level information is available, is that our strategy allowed to balance community-level variables even when these were not included in the matching set in the estimation of the propensity score model. This indicates that the predicted random effects "substitute", in a sense, the cluster-level variables and, hence, achieving a good balance for the former ensures a good balance also for the latter.

References

Aassve, A., Betti, G., Mazzuco, S., Mencarini, L. (2007) Marital disruption and economic well-being: a comparative analysis. *Journal of the Royal Statistical Society: Series A*, 170(3), 781–799.

Aassve, A,, Engelhardt, H., Francavilla, F., Kedir, A., Kim, J., Mealli, F., Mencarini, L., Pudney S. and Prskawetz A. (2006a) Fertility and poverty in developing countries: a comparative analysis. Population Review 45(2).

Aassve. A., Kedir, A. M., Tadesse Weldegebriel, H. (2006b) "State Dependence and Causal Feedback of Poverty and Fertility in Ethiopia", ISER Working Paper No 2006-30.

Abadie, A. (2003) Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113, 231–263.

Abadie, A., Drukker, D., Leber Herr, J. and Imbens, G.W. (2004), Implementing Matching Estimators for Average Treatment Effects in Stata. *Stata Journal*, 4(3), 290-311.

Abadie, A. and Imbens, G. W. (2002) Simple and Bias-Corrected Matching Estimators for Average Treatment Effects. Technical Working Paper T0283, NBER.

Abadie, A. and Imbens. G. W. (2004) On the Failure of the Bootstrap for Matching Estimators. NBER Technical Working Paper T0325.

Abadie, A. and Imbens, G. W. (2006), Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1), 235-267.

Admassie, A. (2002) Explaining the High Incidence of Child Labour in Sub-Saharan Africa. *African Development Review,* 14(2): 251 – 275.

Aitkin, M., Longford, N. (1986) Statistical modelling in school effectiveness studies (with discussion). *Journal of Royal Statistical Society A*, 149, 1-42.

Ali, I. and Pernia, M. E. (2003) Infrastructure and poverty reduction – what is the connection? Economics and Research Department, Policy Brief no. 13, ADB, Manila.

Angrist, J. D. (1998) Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica*, 66(2), 249-288.

Angrist, J. D. and Evans, W. N. (1998) Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*. 88(3), 450–77.

Angrist, J. D. and Imbens G. W. (1991) Sources of Identifying Information In Evaluation Models, Technical Working Paper 117, NBER.

Angrist, J. and Imbens, G. W. (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of American Statistical Association*, 90, 431–442.

Angrist, J. *D*., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association,* 91, 444–472.

Angrist, J. and Krueger, A. (2001) Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *The Journal of Economic Perspectives*, 15(4), 69-85.

Anh, *D*. N. and Thang, N. M. (2002) *Accessibility and Use of Contraceptives in Vietnam*. *International Family Planning Perspectives*, 28(4), 214-219.

Athey, S. and Imbens G. W. (2006) Identification and Inference in Nonlinear Difference-in-Differences Models. Econometrica, 74(2), 431-497.

Augurzky, B. and C. Schmidt (2001) The Propensity Score: A Means to An End. Discussion Paper No. 271, IZA.

Balisacan, A.M., Pernia, E.M. and Estrada, G.E.B. (2003) Economic Growth and Poverty Reduction in Vietnam. In: Pernia, E.M. and Deolalikar, A.B. (Eds) *Poverty, Growth and Institutions in Developing Asia*. Hampshire, England: Palgrave Macmillan Publishers.

Baltagi, B.H. (2001) *Econometric Analysis of Panel Data*, second edition. Chichester: John Wiley & Sons.

Barro, R. J. and Becker, G..S. (1989, Fertility choice in a model of economic growth. *Econometrica*, 57 (2), 481-501.

Becker, G. S., Glaeser, E. L., Murphy, K. M. (1999) Population and economic growth. *The American Economic Review*, 89 (2), 145-149.

Becker, G. S. and Lewis, H.G. (1973) On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81(2), S279-S288.

Becker, S.O. and Ichino, A. (2002) Estimation of average treatment effects based on propensity scores. *The STATA Journal*, 2, 358–377.

Belanger, *D*. (2002) Son preference in a rural village in North Vietnam. *Studies in Family Planning, 33*(4), 321–334.

Black, D. and Smith J. (2004) How Robust is the Evidence on the Effects of the College Quality? Evidence from Matching. *Journal of Econometrics*, 121(1), 99-124.

Blundell, R., Dearden, L. and Sianesi B. (2005) Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS. *Journal of the Royal Statistical Society, Series A*, 168(3), 473-512.

Blundell, R. and Powell, J. (2003) Endogeneity in nonparametric and semiparametric regression models. In: Hansen, L., Dewatripont, M. and Turnovsky, S.J. (Eds.), *Advances in Economics and Econometrics*. Cambridge University Press, Cambridge, pp. 312–357.

Bruno, M., Ravallion, M., Squire, L. (1999) Equity and Growth in Developing Countries. Old and New Perspectives on the Policy Issues, World Bank Policy Research, Working Paper No. 1563.

Bryson, A., Dorsett, R. and Purdon S. (2002) The Use of Propensity Score Matching in the Evaluation of Labour Market Policies. Working Paper No. 4, Department for Work and Pensions.

Caliendo, M. and Kopeining, S. (2005) *Some Practical Guidance for the Implementation of Propensity Score Matching*. IZA working paper, 1588.

Card, D. (1995) Using geographic variation in college proximity to estimate the return to schooling. In: Christofides, L., Grant, E., Swidinsky, R. (Eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press, Toronto, pp. 201–222.

Carneiro, P., Heckman, J. J. and Vytlacil, E. (2005) Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education. Unpublished manuscript, University of Chicago, Department of Economics.

Casterline, J. (1985) *Community effects on fertility* In J.B. Casterline (ed.), The Collection and Analysis of Community Data, Voorburg: International Statistical Institute.

Chun, H. and Oh, J. (2002) An instrumental variable estimate of the effect of fertility on the labour force participation of married women. *Applied Economics Letters*, 9, 631-634.

Coudouel, A., Hentschel, J. and Wodon, Q. (2002) *Poverty Measurement and Analysis*, Poverty Reduction Strategy Paper Sourcebook, World Bank, Washington *D*.C.

Cox, *D*. R. (1958) *Planning of experiment*. New York, Wiley.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A., (2006) Nonparametric Tests for Treatment Effect Heterogeneity. IZA Discussion Paper No. 2091, forthcoming on *Review of Economics and Statistics.*

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A., (2007) Dealing with Limited Overlap in Estimation of Average Treatment Effects, NBER Technical Working Paper No. 330.

Cuong, N. V. (2007) Impact Evaluation of Multiple Overlapping Programs Under a Conditional Independence Assumption. Working Paper N. 27 Mansholt Graduate School. Available at http://www.sls.wau.nl/mi/mgs/publications/ Mansholt_Working_Papers/MWP_36.pdf

Das, M. (2005) Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics* 124, 335–361.

Davis, K. and Blake. J. (1956) Social structure and fertility. *Economic Development and Cultural Change*, 4(3), 211-235.

Davis, J. A., Spaeth, J. L. and Huson, C. (1961) A technique for analyzing the effect of group composition. *American Sociological Review*, 26, 215-225.

Dawid, A. P. (1979) Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society B*, 41, 1-31.

Dawid, A.P. (2002) Influence Diagrams for Causal Modeling and Inference. *International Statistical Review*, 70, 161-189.

Deaton, A. and Zaidi, S. (2002), Guidelines for Constructing Consumption Aggregates for Welfare Analysis, Living Standards Measurement Study Working Paper No. 135, The World Bank.

Dehejia, R., and Wahba, S. (1999) Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94, 448, 1053–1062.

Diez-Roux, A. V. (2004) Estimating neighborhood health effects: the challenges of causal inference in a complex world. *Social Science & Medicine*, 58(10), 1953-1960.

Di Prete, T.A. and Forristal, J. D. (1994) Multilevel Models: Methods and Substance, *Annual Review of Sociology*, 20, 331-357.

Drovandi, S. and Salvini. S. (2004) Women's Autonomy and Demographic Behavior. *Population Review,* 43(2).

Durbin, J. (1954), Errors in Variables, *Review of the International Statistical Institute*, 22, 23-32.

Durkheim, E. (1897) [1951] *Suicide*, Glencoe, III: Free Press.

Duy, L. V., Haughton *D*., Haughton J., Kiem *D*. A. and Ky L. *D*. (2001) Fertility decline. In: Haughton *D*., Haughton J., Phong N. (eds), *Living Standards during an Economic Boom. Vietnam 1993-1998*, Statistical Publishing House, Hanoi.

Easterlin, R. A. (1975) An Economic Framework for Fertility Analysis, *Studies in Family Planning*, 6(3), 54-63.

Easterlin, R.A. and Crimmins E.M. (1985) *The Fertility Revolution*. Chicago: University of Chicago Press.

Ebbes, P., Bockenholt, U. and Wedel, M. (2004) Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58, 161–178.

Entwisle, B., Casterline, J. B. and Hussein, A-A. S. (1989) Villages as Contexts for Contraceptive Behavior in Rural Egypt. *American Sociological Review,* 54,1019-1034.

Eren, O. (2007) Measuring the Union-Nonunion Wage Gap Using Propensity Score Matching. *Industrial Relations*. 46(4).

Evans, M., Gough, I., Harkness, S., McKay, A., Thanh, H. D. and Le Thu, N. D. (2007) How Progressive is Social Security in Viet Nam? UNDP Vietnam Policy Dialogue Paper, available at http://www.undp.org.vn/undpLive/digitalAssets/7589_SS_Progressive__E_.pdf

Falaris, E.M. (2003) The effect of survey attrition in longitudinal surveys: evidence from Peru, Cote d'Ivoire, and Vietnam. *Journal of Development Economics*, 70, 133-157.

Fielding, A. (2004) The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed. *Multilevel Modelling Newsletter*, 16(2), 3–9.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*. 1st Edition. Oliver and Boyd, Edinburgh.

Florens, J., Heckman, J., Meghir, C. and Vytlacil, E. (2002) Instrumental variables, local instrumental variables and control functions. Cemmap Working Paper 15/02.

Frangakis, C. E., and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, 58, 21-29.

Frolich, M. (2007) Non parametric IV estimation of local average treatment effects with covariates, *Journal of Econometrics*, 139, 35-75.

Garza-Rodríguez, J. (2004) The Determinants of Poverty in México: 2002. *Proceedings of the 8th International Conference on Global Business and Economic Development*. Guadalajara, Jalisco, México.

Ghura, D., Leite, C. A., Tsangarides, C. G. (2002) Is growth enough? Macroeconomic Policy and Poverty Reduction, IMF (International Monetary Fund) Working paper, Wp/03/118.

Gitelman, A. I. (2005) Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, 30, 397-412.

Glewwe, P., Gragnolati, M. and Zaman, H. (2002) Who Gained from Vietnam's Boom in the 1990s? *Economic Development and Cultural Change*, 50(4), 773-792.

Goldberger, A. (1972) Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6), 979-1001.

Goldstein, H. (1995) *Multilevel Statistical Models*, Edward Arnold, London.

Goldstein, H., Browne, W. J and Rasbash, J. (2002) Partitioning variation in multilevel models, *Understanding Statistics,* 1, 223-232.

Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, A*, 158, 175-177.

Goldstein H., Thomas S. (1996) Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society A,* 159, 149-63.

Goodman, A. and Sianesi, B. (2005) Early Education and Children's Outcomes: How long the impacts last. *Fiscal Studies*, 26(4).

Green, W.H. (2002) *Econometric Analysis,* 5th edition, Prentice Hall.

Greenland, S. and Brumback, B. (2002) An overview of relations among causal modeling methods. *International Journal of Epidemiology*, 31, 1030-1037.

Grilli, L. and Rampichini, C. (2006) Model building issues in multilevel linear models with endogenous covariates. Paper presented at the "KNEMO: Knowledge Extraction and Modeling" IASC-INTERFACE-IFCS Workshop, September, 4th-6th 2006 Villa Orlandi Island of Capri, ITALY.

GSO (General Statistical Office) (1994) *Vietnam Living Standards Survey - 1992/93. Basic iformation*, Hanoi.

GSO (General Statistical Office) (2000) *Vietnam Living Standards Survey - 1997/98. Basic information*, Hanoi

Gupta, N. D. and Dubey, A. (2003) Poverty and Fertility: An Instrumental Variables Analysis on Indian Micro Data, Working Paper 03-11, Aarhus School of Business.

Haavelmo, T. (1943) The statistical implication of a system of simultaneous equations. *Econometrica*, 11, 1–12.

Hahn, J. (1998) On the role of the propensity scores in effcient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315-331.

Hahn, J. Todd,P. and Van der Klaauw, W. (2001) Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica,* 69(1), 201-209.

Halloran, M. E. and Struchiner, C. J. (1995) Causal inference in infectious disease. *Epidemiology,* 6, 142-151.

Harding, D. J. (2003) Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology*, 109, 676–719.

Hardle, W. and Linton, O. (1994) Applied Nonparametric methods, in *Handbook of Econometrics*, Volume 4, ed. Engle, R. F. and McFadden, D., Amsterdam: North Holland, 391-448.

Hartwell, R. M. (1986) The long debate on poverty. Paper presented at the Political Economy Seminar Series, St. Louis, Missouri: Washington University.

Haughton, J. and Haughton. D. (1995) Son preference in Vietnam. *Studies in Family Planning*, 26 (6), 325-337.

Haughton, D., Haughton, J. and Phong, N. (edited by) (2001) *Living Standards during an Economic Boom. Vietnam 1993-1998*, Statistical Publishing House, Hanoi.

Hausman, J. A. (1978) Specification tests in econometrics. *Econometrica*, 46, 1251-1272.

Heckman, J. J. (1979) Sample selection bias as a specification error. *Econometrica*, 47, 1, 153-161.

Heckman, J. J. (1992) Randomization and social program evaluation, In: Masnski, C. F. and Garfinkel, I. (Eds.), *Evaluating welfare and training programs,* Cambridge, MA: Harward University Press, 201-230.

Heckman, J. J. (1996) Comment on "Identification of causal effects using instrumental variables" by Angrist J. D., Imbens, G. W., and Rubin D. B. *Journal of the American Statistical Association*, 91, 459–462.

Heckman, J. J. (1997) Instrumental Variables: A study of implicit behavioural assumptions used in making program evaluations. *Journal of Human Resources*, 32, 441-462.

Heckman, J.J., Hotz V.J. (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, 408, 862-874.

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

Heckman, J.J., Ichimura, H. and Todd, P. (1998) Matching as an Econometric Evaluation Estimator. *Review of Economic Studies*, 65(2), 261-294.

Hirano, K., Imbens, G. and Ridder G. (2003), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, *Econometrica*, Vol. 71(4).

Hirano, K., Imbens, G., Rubin, *D*. and Zhou, X. (2000) Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1, 69–88.

Hirschman, C. and Guest, P. (1990) Multilevel models of fertility determination in four Southeast Asian countries: 1979 and 1980. *Demography*, 27 (3), 369-396

Holland, P. (1986) Statistics and causal inference. *Journal of American Statistical Association*, 81, 945–970.

Hong, G. (2003) Causal Inference for Multilevel Observational Data with Application to the Kindergarten Retention Study, paper presented at the Joint Statistical Meetings, San Francisco, 2003.

Hong, G., and Raudenbush, S. W. (2005) Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27(3), 205-224.

Hong, G., and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, 101(475), 901-910.

Hox, J.J. (1995) *Applied Multilevel Analysis*. Amsterdam: TT- Publikaties.

Huong, P. L., Tuan, B. Q. and Minh, D. H. (2003) Employment Poverty Linkages and Policies for Pro-Poor Growth in Viet Nam, ILO Issues in Employment and Poverty, Discussion Paper 9, Geneva.

Hwang, H. S. (1980) A Comparison of Tests of Overidentifying Restrictions. *Econometrica*, 48(7), 1821-1825.

Ichino, A., Mealli, F. and Nannicini, T. (2006), From Temporary Help Jobs to Permanent Employment: What Can We Learn from Matching Estimators and their Sensitivity*?*, CEPR Discussion Paper N.5736.

Ichino, A., Mealli, F. and Nannicini, T. (2007) From Temporary Help Jobs to Permanent Employment: What Can We Learn from Matching Estimators and their Sensitivity? *Journal of Applied Econometrics*, forthcoming.

Imai, K, and van Dyk, D. A. (2003) Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99, 854-866

Imbens, G. W. (2000) The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika,* 87(3), 706-710.

Imbens, G. W. (2002) A Classification of Assignment Mechanisms. Mimeo, available at http://elsa.berkeley.edu/users/imbens/e244_f02/chap3.pdf.

Imbens, G. W. (2003) Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review*, 93(2), 126-132.

Imbens, G. W. (2004) Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review, *Review of Economics and Statistics*, 86, 4-30.

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

Imbens, G. W. and Hirano, K., (2004) The Propensity score with continuous treatment, chapter for Missing data and Bayesian Method in Practice: Contributions by Donald Rubin Statistical Family.

Johansson, A. (1996) Family planning in Vietnam - women's experiences and dilemma: a community study from the Red River Delta. *Journal of Psychosomatic Obstetrics and Gynecology*, 17, 59-67.

Johansson, A. (1998) Population policy, son preference and the use of IUDs in North Vietnam. *Reproductive Health Matters*, 6, 66-76.

Josipovič, D. (2003) Geographical factors of fertility, *Acta Geographica Slovenica*, 43, 111-118.

Justino, P. (2005) Beyond HEPR: A Framework For An Integrated National System Of Social Security In Viet Nam. UNDP Vietnam Policy Dialogue Paper 2005/1.

Justino, P. and Litchfield, J. (2004) Welfare in Vietnam During the 1990s: Poverty, Inequality and Poverty Dynamics. *Journal of the Asian Pacific Economy*, 9(2), 145-169.

Kabeer, N. (2001) Deprivation, discrimination and delivery: competing explanations for child labor and educational failure in South Asia. Institute of Development Studies Working Paper 135, Sussex, Brighton, UK

Kim, J. and Aassve A. (2006) Fertility and its Consequence on Family Labour Supply, Institute of Labour Studies (IZA) Discussion Paper No. 2162.

Kim, J., Engelhardt, H., Prskawetz, A., and Aassve, A. (2005) Does fertility decrease the welfare of households? An analysis of poverty dynamics and fertility in Indonesia. Working Paper 5, Vienna Institute of Demography.

Kim, J. and Seltzer, M. (2007), Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles.

Klepinger, D., Lundberg, S. and Plotnick, R. (1995) Instrumental selection: the case of teenage childbearing and women's educational attainment. Unpublished manuscript, University of Washington.

Kreft, I. G. and de Leeuw, J. (1998) *Introducing Multilevel Modelling*. Sage, London.

Kreft, I. G., de Leeuw, J. and Aiken, L. (1995) The effect of different forms of centering in hierarchical linear models. *Journal of educational statistics*, 2, 171 – 186.

Lanjouw, P. and Ravallion, M. (1995) Poverty and household size. *Economic Journal*, 105, 433, 1415-1434.

Lechner, M. (2000) An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany. *Journal of Human Resources*, 35(2), 347-375.

Lechner, M. (2001) Identification And Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In Lechner, M. and Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Heidelberg: Physica-Verlag.

Lee, W. (2006) Propensity Score Matching and Variations on the Balancing Test," Unpublished Manuscript available at http://melbourneinstitute.com/ people/wlee/Balancing_Test_Paper.pdf.

Livi-Bacci, M. (1994) *Poverty and population*, [reprint] IUSSP Distinguished lecture series on population and development, International Conference on Population and Development (ICPD), reprints, Li.

Livi-Bacci, M. (2000) *A concise history of the world population*. Blackwell: Oxford.

Livi-Bacci, M. and De Santis, G. (1998) *Population and Poverty in the Developing World*. Clarendon Press: Oxford.

Maas, C.J.M. and Hox, J.J. (2004) Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2), 127–137.

Macintyre, S. (2000) The social patterning of health. Bringing the social context back in. *Medical Sociology Newsletter,* 26, 14-19.

Maddala, G. S. (1983) *Limited dependent and qualitative variables in econometrics.* Cambridge, U.K., Cambridge University Press.

Maldonado, G. and Greenland, S. (2002) Estimating causal effects. *International Journal of Epidemiology*, 31, 422-38.

Malthus, T.R., (1798) *An essay on the principle of population*. Reeves and Turner: London.

Manski C.F. (1990), Nonparametric Bounds on Treatment Effects, *American Economic Review Papers and Proceedings*, 80, 319-323

Manski, C. F. and Garfinkel, I. (1992) *Evaluating Welfare and Training Programs*, Harvard University Press: Cambridge.

Marx, K. (1846) [1957] The German ideology, New York: Int. Publ.

McClellan, M., McNeil, B. J. and Newhouse, J. P. (1994) Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? *Journal of the American Medical Association*. 272(11), 859 – 866.

McNicoll, G. (1997) Population and poverty: A review and restatement. Policy Research Division Working Paper No. 105, New York: Population Council.

Moav, O. (2005) Cheap Children and the Persistence of Poverty. *The Economic Journal,* 115, 88-110.

Moffitt, R. A. (1996) Comment on "Identification of causal effects using instrumental variables" by Angrist J. D., Imbens, G. W., and Rubin D. B. *Journal of the American Statistical Association*, 91, 462–465.

Molini, V. (2006) Food Security in Vietnam during the 1990s - The Empirical Evidence. United Nation University, Reaserch paper No. 2006/67.

Mukherjee, S. and Benson, T. (2003) The Determinants of Poverty in Malawi, 1998. *World Development*, 31(2), 339 – 358.

Mundlak, Y. (1978) On pooling of time series and cross section data. *Econometrica,* 64, 69–85.

Muthén, B. O. (2002) Beyond SEM: General Latent Variable Modeling. *Behaviormetrika*, 29, 81-117.

Nannicini, T. (2007) Simulation-Based Sensitivity Analysis for Matching Estimators. *The Stata Journal*, 7(3), 334-350.

Newey, W., Powell, J. and Vella, F., (1999) Nonparametric estimation of triangular simultaneous equations models. *Econometrica,,* 67, 565–603.

Neyman, J. (1923) On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in *Statistical Science*, 5(4), 465–480, (1990).

Niimi, Y., Vasudeva-Dutta, P. and Winters, A. (2003) Trade Liberalization and Poverty Dynamics in Vietnam, PRUS Working Paper, No. 17, Poverty Research Unit at Sussex, University of Sussex.

Oakes, J. M. (2004) The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. *Social Science and Medicine*, 58, 1929–1952.

Oehlert, G. W. (1992) A note on the delta method. *American Statistician*, 46, 27–29.

Pearl, J. (1995) Causal Diagrams for Empirical Research (with Discussion). *Biometrika*, 82, 669-710.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Puhani, P. (2000) The Heckman Correction for Sample Selection and Its Critique - A Short Survey. *Journal of Economic Surveys*, 14, 53–68.

Pudney, S. and Aassve, A. (2007a) Poverty transitions in developing countries: the roles of economic and demographic change, *Working Paper of Institute for Social and Economic Research*, paper 2007-25. Colchester: University of Essex

Pudney, S. and Aassve A. (2007b) Endogenous fertility and its impact on poverty: Evidence from Vietnam, *Working Paper of Institute for Social and Economic Research*, paper 2007-26. Colchester: University of Essex

Rabe-Hesketh, S. and Skrondal, A. (2005) *Multilevel and Longitudinal Modeling Using Stata*. Stata Press books, StataCorp LP.

Ravallion, M. and Bidani, B. (1994) How robust is a poverty profile? *World Bank Economic Review*, 8(1), 75-101.

Reiersøl, O. (1945) Confluence Analysis by Means of Instrumental Sets of Variables. *Akiv för Matematik*, *Astronomi och Fysik*, 32a, 1-119

Rice, N., Jones, A.M. and Goldstein, H. (2002) Multilevel models where the random effects are correlated with the fixed predictors, Manuscript available at: http://www.mlwin.com/hgpersonal/cigls.pdf.

Roberts, S. (1999) Socioeconomic composition and health: The independent contribution of community socioeconomic context. *Annua1 Review of Sociology,* 25, 489-516.

Robins, J. M., and Rotnitzky, A. (1995), Semiparametric Efficiency in Multivariate Regression Models with Missing Data, *Journal of the American Statistical Association*, 90, 122-129

Robinson, W. S. (1950) Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Rosenbaum, P. R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, 147, 656–666.

Rosenbaum, P. R. (1987) The Role of a Second Control Group in an Observational Study. *Statistical Science*, 2(3), 292-306.

Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosenbaum P. and Rubin D. (1983b), Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society, Series B*, 45, 212-218.

Rosenbaum, P. R. and Rubin, D. B. (1984) Reducing bias in observational studies using subclassication on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39(1), 33-38.

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

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

Rubin, D., (1980) Discussion of Randomization Analysis of Experimental Data: The Fisher Randomization Test by D.Basu. *Journal of the American Statistical Association*, 75, 591-93.

Schoumaker, B. and Tabutin, D. (1999) *Relations entre pauvreté et fécondité dans les pays du sud. Etat des connaissances, méthodologie et illustrations*. SPED Document de Travail, No. 2, Feb. 1999, Université Catholique de Louvain, Département des Sciences de la Population et du Développment: Louvain-la-Neuve, Belgium.

Self, S. G. and Liang, K. L. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.

Sen A. K. (1987), *The Standard of Living*, Cambridge: Cambridge University Press.

Simon, H. (1954) Spurious Correlation: A Causal Interpretation. *Journal of the American Statistical Association*, 49, 467-479.

Skinner, B. F. (1965) *Science and human behaviour*, New York: Free Press.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL: Chapman & Hall/CRC.

Small, D. (2007) Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102, 1049-1058.

Small, D. and Rosenbaum, P. R. (2007) War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. Accepted for publication in the *Journal of the American Statistical Association.*

Smith, J. (2000) A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies. *Schweizerische Zeitschrift fuer Volkswirtschaft und Statistik*, 136(3), 1-22.

Smith, J. and P. Todd (2005) Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*. 125(1-2), 305-353.

Snijders, T. A. B. and Berkhof, J. (2007) Diagnostic checks for multilevel models. In: Jan de Leeuw and Erik Meijer (Eds.), *Handbook of Multilevel Analysis*, New York: Springer.

Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*, London: Sage.

Sobel, M. E. (2006) Spatial Concentration and Social Stratification: Does the Clustering of Disadvantage 'Beget' Bad Outcomes? In S. Bowles, S.N. Durlauf, and K. Hoff (Eds.), Poverty Traps, pp. 204-229, New York: Russell Sage Foundation.

Spencer, N. H. And Fielding, A. (2000) An instrumental variable consistent estimation procedure to overcome the problem of endogenous variables in multilevel models. Multilevel Modelling Newsletter, 12(1), 4-7.

Spirtes, P., Glymour, C. and Scheines R. (1993) *Causation, Prediction and Search*. Berlin: Springer-Verlag.

S.S.A. (United States Social Security Administration) (2007) *Social Security Programs Throughout the World: Asia and the Pacific, 2006* (released March 2007). Available at http://www.ssa.gov/ policy/docs/progdesc/ssptw/2006-2007/asia/vietnam.pdf

STATA (2005) *Longitudinal/Panel Data Reference Manual*. Texas, USA: Stata Press.

Stock, J. and Trebbi, F. (2003) Who Invented Instrumental Variable Regression? *Journal of Economic Perspectives*, 17, 177-194.

Stuart, E. A. (2007 Estimating causal effects using school-level datasets. *Educational Researcher*, 36, 187-198

Subramanian, S. V. (2004) The relevance of multilevel statistical methods for identifying causal neighbourhood effects. *Social Science & Medicine*, 58(10), 1961-1967.

Subramanian, S. V., Jones, K. and Duncan, C. (2003) *Multilevel methods for public health research*. In: Kawachi, I. and Berkman, LF. (Eds). *Neighborhoods and Health*, New York: Oxford University Press, pp. 65-111.

Subramanian, S. V., Kawachi, I. and Kennedy, B. P. (2001) Does the state you live in make a difference? Multilevel analysis of self-rated health in the US. *Social Science and Medicine*, 53, 9-19.

Testa, M. R. and Grilli, L. (2006) The influence of childbearing regional contexts on ideal family size in Europe: A multilevel analysis. *Population*, 61(1-2), 107-137.

Tung, P. D. (2004) *Poverty line, poverty measurement, monitoring and assessment of MDG in Viet Nam*. Report presented at the "2004 International Conference on Offical Poverty Statistics – Methodology and Comparability", Manila, October 2004.

Van de Walle, D. (1996) Infrastructure and Poverty in Vietnam, Living Standards Measurement Study Working Paper No. 121, The World Bank Group, Washington DC.

Van de Walle, D. and Gunewardana, D. (2001) Sources of ethnic inequality in Vietnam. *Journal of Development Economics*, 65, 177-207.

Wald, A. (1940) The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11, 284-300.

Weber, M. (1905) [1958] The protestant ethic and the spirit of capitalism, New York: Scribners.

White, H. and Chalak, K. (2006) A Unified Framework for Defining and Identifying Causal Effects. UCSD Deptartment of Economics Discussion Paper.

White, H. and Masset, E. (2002) Child poverty in Vietnam: using adult equivalence scales to estimate income-poverty for different age groups, MPRA Paper 777, University Library of Munich, Germany.

White, H. and Masset, E. (2003) Constructing the Poverty Profile: An Illustration of the Importance of Allowing for Household Size and Composition in the Case of Vietnam, Young Lives Working Paper No. 3, London: Young Lives and Save the Children Fund UK.

Willis, R. J. (1973) A New Approach to the Economic Theory of Fertility Behavior. *Journal of Political Economy,* 81(2), part 2: S14-S64.

Wooldridge, J.M. (2002) *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

World Bank (2000) *Vietnam: Attacking Poverty - Vietnam Development Report 2000,* Joint Report of the Government-Donor-NGO Working Group, Hanoi.

World Bank (2005), Introduction to Poverty Analysis, Manuscript available at http://info.worldbank.org/etools/library/latestversion.asp?207005.

Wright, P. G. (1928) *The Tariff on Animal and Vegetable Oils.* New York: Macmillan.

Zhao, *Z*. (2005) Sensitivity of Propensity Score Methods to the Specifications. IZA Discussion Paper No. 1873