The effect of fertility on poverty: an example of causal inference with multilevel data in demographic research

Bruno Arpino

Department of Decision Sciences and "Carlo F. Dondena" Center for Research on Social Dynamics, Bocconi University, Via Roentgen 1, 20136 Milano, Italy,
e-mail: bruno.arpino@unibocconi.it

PRELIMINARY DRAFT – PLEASE DO NOT QUOTE OR CITE

Abstract Multilevel structured data are very common in demographic research: e.g., individual clustered in households; households clustered in communities. In this paper we focus on the role of the community context in the estimation of the causal effect of fertility on poverty in Vietnam. The contextual characteristics can strongly influence both poverty and fertility and therefore it is crucially important to account for this in order to draw valid causal inference. The multilevel dimension introduces statistical complications and stimulates interesting research questions. From the methodological point of view, we use multilevel techniques in the propensity score matching implementation and a weaker version of the traditional SUTVA assumption. We find a negative and substantial effect of fertility on wellbeing; this effect is stronger in high-level fertility communities. On the contrary, fertility measured at the community level does not have a significant effect per se on the individual wellbeing.

Keywords: fertility-poverty relationship, causal inference, multilevel data, propensity score matching, SUTVA.

1. Introduction

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 topic was studied at the national or state level. In the last decades, the micro approach, which usually takes the household as unit of analysis, has been remarkably developed. 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 large number 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.

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.

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 economics (e.g. Rosenmbaum and Rubin, 1983; Heckman, 1992 and 1997a; Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Heckman, Ichimura and Todd, 1997) and has started to be widely employed also in demography (e.g., Aassve et al, 2007; Engelhardt et al, 2009).

Adopting the potential outcome framework, we use recorded childbearing events between the two waves as a measure of fertility and constitute our treatment variable. The outcome variable is defined as the variation in household consumption expenditures between the two waves. Consequently, each household has two potential outcomes: the consumption expenditure variation Y_1 if it experiences a childbearing event between the two waves (treated) and Y_0 otherwise (untreated or control). But childbearing is, at least in part, down to individual choice, giving rise to *self-selection*: households that choose to have more children (self selected into the treatment) may

be very different from households that choose to have fewer children irrespective of the treatment. Hence, if we observe that the first group of households has on average lower per capita expenditure, we cannot necessarily assert that this is due to fertility since the two groups of households are likely to be different in respect to many other characteristics, such as education. Thus, a simple difference in the average consumption (or income) for the two groups of households gives a biased estimate.

In principle, we would compare units of similar characteristics that differ only by the treatment status. For these units the observed difference in the outcome can be reasonably assumed to be due to the treatment. Propensity score matching (PSM) relies on the selection on observables assumption, which is referred to as the Unconfoundedness Assumption (UNC). Multiple regression is also a method relying on this assumption, though the identifying assumption can be stated in a weaker way (see e.g. Wooldridge, 2002). However, regression analyses impose additional assumptions, like linearity in the treatment effect, that can be overcome by using a matching approach.

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 both their poverty status and fertility behaviour. 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 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.

We stress that the methodological and substantive issues connected to the multilevel data structure we have in our work are not peculiar to this application but are relevant in virtually all studies of causal inference in social fields and demography in particular. In fact, multilevel structures are omnipresent are often relevant and interesting to analyse.

The paper is organised as follow. In section 2 we briefly review the literature on the fertility-poverty relationship, present the Vietnamese context and data. In section 2 we introduce the potential outcomes framework and discuss the statistical complications arising in a multilevel setting. Section 4 shows the results and section 5 concludes.

2. Background

2.1. The literature on the relationship between fertility and poverty

In the last decades the micro approach to the study of the fertility-poverty relationship has been remarkably developed. 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 (Livi-Bacci and De 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, 2006).

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 and allow to properly implement a pre-post treatment study which is vital for our study of causal inference.

2.2. The Vietnam Living Standard Measurement Survey and the Vietnamese context

We use data 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. 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.

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 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. 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 multiple 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.

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) which testimony a substantial poverty reduction in the survey period.

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. Much of this improvement has been attributed to the "Doi Moi" policy (translated in English as "renovation") that was initiated in the late 1980s with many similarities with the reforms taking place in China a decade earlier.

During the nineties, immediately following the Doi Moi, Vietnam experienced a positive macroeconomic trend and 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.

Whereas the economic boom in Vietnam affected all geographical, ethnic, and socio-economic groups, the poverty reduction was certainly not uniform across the population (Justino and Litchfield, 2004; Balisacan et al, 2003; Glewwe et al, 2002). Gains from economic growth was stronger in urban areas, for South East and Red River Delta[1] for Kinh which is the main ethnic group in Vietnam, for households headed by a white collar worker and for those with higher education. However, empirical evidence also shows much stronger heterogeneity in poverty reduction in rural areas. There is in other words a significant degree of clustering across rural areas and this is one of the reason 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

[1] The Red River Delta and the Mekong River Delta were the regions that benefited more from rice market liberalisation (Justino and Litchfield, 2004).

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 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. 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 role.

2.3. Why adopting a multilevel perspective in studying the relationship between fertility and poverty?

In general, it is not possible to 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 both for poverty conditions and fertility behaviours.

In addition to the individual and household characteristics, the literature on poverty analysis has been placing an increasing focus on the role 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).

Several geographical and institutional characteristics impact considerably on poverty. In general, poverty is high in areas characterized by geographical isolation, poor resources, 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 system. Regional-level market reforms can boost growth and help poor people.

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.

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.

As happened for poverty, also in the literature on fertility studies, the role of the contextual characteristics 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, presence of family planning centres, 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.

The fact that the determinant of fertility and poverty are individual, household and contextual characteristics motivates the multilevel perspective we adopt in the paper. 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. Failing to recognize the hierarchical nature of the data can have important statistical consequences on our study of causal inference and can hide important research questions. First of all, we need to recognize that community level characteristics are potentially important confounders to control for in our analysis. For the rural sample the VLSMS includes, as aforesaid, essential data on community characteristics allowing to bring an important part of information into the analysis.

Moreover, the multilevel dimension implies specific statistical challenges for causal inference that we discuss extensively in section 3. However, 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 produce 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.

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 sources of heterogeneity can be partly related to the overall level of fertility in the community. This violates one of the assumptions usually invoked in the potential outcome literature, namely the SUTVA. In section 3.2 we discuss a simple approach to overcome this problem, highlighting interesting causal quantities that are natural to estimate as a consequence of the proposed approach.

3. Methodology

3.1. Causal inference in observational studies using the potential outcomes approach

To formalise the idea of the potential outcomes approach (Rubin 1974 and 1978), suppose we have a sample of individual units under study indexed by i = 1, 2, ... N, a treatment indicator D that assumes the value 1 for treated units and 0 for untreated or the controls and an outcome variable, here indicated by Y. An important assumption employed in this setting is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980), which states that the potential outcomes for any unit are not influenced by the treatments assigned to any other units and that there are no hidden versions of the treatment. Under the SUTVA 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. X indicates the set of covariates influencing both potential outcomes and the selection into the treatment – also known as the confounding variables. The two causal parameters of interest are the *Average Treatment Effect* (ATE) and the *Average Treatment Effect on the Treated* (ATT) which are defined as:

$$\text{ATE} = E(Y_{i1} - Y_{i0}) \tag{1}$$

$$\text{ATT} = E(Y_{i1} - Y_{i0} | D_i=1). \tag{2}$$

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). The most prominent evaluation parameter is usually 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 key issue for causal inference in observational studies is that units are not necessarily assigned randomly to the treatment and the control groups. On the contrary, the observed individuals decide (at least to some extent) to take one of the two treatments which tends to depend on characteristics also influencing the outcome of interest, and hence confounding the causal relationship. Consequently, simply comparing the outcome of treated and controls give biased estimate of the causal effects. This is the well known problem of *self selection bias*.

The assumption that selection takes place only on observable characteristics is also known as the unconfoundedness assumption (UNC) and represents the fundamental identifying assumption in several empirical works[2]:

Assumption A.1 (Unconfoundedness)

$$Y_1, Y_0 \perp D \mid X$$

where $\perp$ in the notation introduced by Dawid (1979) indicates independence. Assumption A.1 implies that after conditioning on variables influencing both the selection and the outcome, the dependence among potential outcomes and the treatment is cancelled out. Regression and matching techniques, as well as stratification and weighting methods, all rely on this assumption. In the regression analysis it usually suffice to assume that conditional independence of potential outcomes on the treatment hold in expected values (see e.g., Wooldridge, 2002, p. 607). That is, we could substitute assumption A.1 with the weaker: $E(Y_1 \mid D, \mathrm{X}) = E(Y_1 \mid \mathrm{X})$ and $E(Y_0 \mid D, \mathrm{X}) = E(Y_0 \mid \mathrm{X})$. The fundamental idea behind these methods is to compare treated units with control units that are similar in their characteristics. Another assumption, termed *overlap*, is also required:

Assumption A.2 (Overlap)

$$0 < P(D=1|X) < 1.$$

where $P(D=1|X)$ is the conditional probability of receiving the treatment given covariates, X. Assumption A.2 implies equality in the support of X in the two groups of treated and controls (i.e. Support($X|D$=1) = Support($X|D$=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 there are no comparable units.

[2] The unconfoundedness assumption has been referred to also as the conditional independence or the exogeneity assumption (Imbens 2004).

Among the methods based on the UNC assumption we focus on regression and matching estimators. In the standard multivariate regression model assuming a linear relationship between outcome and independent variables and homogeneity of treatment effects, the ATE would coincide with the ATT and we are not able to make separate estimates of the two quantities. Moreover, if the true model was non linear, the OLS estimates of the treatment effects would be in general biased. In parametric regression the overlap assumption is not required in so far we can be sure to have the correct specification of the model. Otherwise the comparison of treated and control units outside the common support rely heavily on the linear extrapolation. Of course, the standard model can be extended and made flexible to overcome these limitations. For example, the common support problem can be circumvented by first estimating it and running the regression conditioning on it. Moreover, heterogeneous treatment effects can be allowed by including a complete set of interactions between X and D. This gives rise to the so-called Fully Interacted Linear Model (FILM in the following – see Goodman and Sianesi 2005). Also the linearity assumption can be removed if we use a non-parametric regression technique. However, non-parametric regression gives rise to the curse of dimensionality as the number of explanatory variables increases. This problem is common to other non-parametric methods, including matching.

A popular way to overcome the dimensionality problem is to implement the matching on the basis of a univariate *Propensity Score* (Rosenbaum and Rubin 1983). This is defined as the conditional probability of receiving a treatment given pre-treatment characteristics: $e(X) \equiv Pr\{D = 1|X\} = E\{D|X\}$. When the propensity scores are balanced across the treatment and control groups, the distribution of all covariates X, are balanced in expectation across the two groups (balancing property of the propensity score). Therefore, matching on the propensity score is equivalent of matching on X. Once the propensity score is estimated, several methods of matching are available. The most common ones are kernel (gaussian and epanechnikov), nearest neighbour, radius and stratification matching (for a discussion about these methods see Caliendo and Kopeinig 2005; Becker and Ichino 2002).

When one or more confounders are not observed assumption A.1. cannot be invoked and matching and regression methods cannot be employed. The most common approach to deal with selection on unobservables is to exploit the availability of a variable assumed to impact the selection into treatment but to have no direct influence on the outcome, which is termed instrumental variable (IV). The concrete possibility to use an IV method relies, of course, on the availability of such a variable. In practice, instruments are often difficult to find. In this case a sensitivity analysis becomes very useful because it can be used to assess the importance of the

violation of the UNC for the estimated causal effect. These issues are discussed by Arpino and Aassve (2008).

3.2. Causal inference in a multilevel setting

In many demographic research analyses data show a multilevel structure. Typically these structures are naturally occurring ones: individuals are grouped in households; households are nested within communities; and communities are clustered in regions. The units in such a system lie at four different levels of a hierarchy. In our application we consider a two-level data structure with households (first level units) grouped in communities (second level units or clusters). Since we are interested in the estimation of causal effects in such a population it is 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.

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. In the paper we focus on the last two issues and refer the interested reader to Arpino (2008) for the cluster-heterogeneity in the estimated causal effect of fertility on poverty.

3.2.1 Addressing the multilevel nature of the selection process

As discussed in section 2.3 contextual factors can be important confounders in the causal relationship of interest together with the individual ones. Multilevel modelling techniques have been specifically implemented to bring together, simultaneously, macro and micro level variables while accounting for the dependence of observations within groups (see e.g., Goldstein, 1995; Snijders and Bosker, 1999). In this paper we explore the use of multilevel models for the estimation of the propensity score.

In a two-level setting, it turns out natural to specify the unconfoundedness assumption as follows:

Assumption A.3 (Unconfoundedness assumption – two-level data)

$$Y_1, Y_0 \perp D \mid X, C$$

where X and C indicates, respectively, first-level and second-level covariates. Under this assumption we can estimate the causal effect of the treatment if we are able to balance all the relevant first and second level characteristics in the treated and control group.

In the propensity score related literature, relatively little attention has been given to the specification of the propensity score when the data have a hierarchical structure, as well as to the underlying identification assumption. For the best of our knowledge only Kim and Seltzer (2007), Aussem (2008), Su (2008), and Li et al (2009) address the issue explicitly. 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, we then achieve automatically a perfect balancing in all the observed and unobserved cluster characteristics. This strategy is likely to be unfeasible in those situations, representing the norm in social and economic observational studies, where we have relatively few units within each cluster. In these cases, in fact, it is likely that in several clusters it is difficult to find for each treated unit a good matched control, with respect to individual characteristics, belonging to the same cluster. This is also our situation since in our application the number of households in a community ranges from 7 to 21.

Excluding strategies that imply a within-cluster matching we propose to take into account the multilevel data structure by including in the propensity score model the cluster-level covariates and/or by using multilevel or fixed effects models. The models we compare are as follows:

Model 1 (simple single level model):

$$\pi_i = F\left(\lambda X_i + \partial C_i\right)$$

Model 2 (two-level random effects model):

$$\pi_{ij} = F\left(\lambda X_{ij} + \partial C_j + u_{0j} + X_{ij}U\right)$$

Model 3 (single level model with clusters indicators - "dummy model"):

$$\pi_{ij} = F\left(X_{ij}\lambda + C_{1j}\,\gamma_1 + C_{2j}\,\gamma_2 + \ldots + C_{Jj}\,\gamma_J\right)$$

where $F(.)$ is the logistic cumulative distribution function and C_1, ..., C_J are indicator variables for clusters.

Multilevel specifications for the propensity score (model 2) should help mitigating the biasing effect of unobserved macro level covariates. We compare these multilevel specifications with single level ones (model 1). We also compare multilevel specifications with alternative approaches which try to take account of unobserved cluster effects. Instead of using a random variable to represent the cluster effects, as in multilevel models, an alternative could be that of estimating fixed intercepts by including a dummy variable for each cluster; we will refer to this approach as the *dummy model* (model 3). In the logistic regression literature, it is well known that this approach can give rise to inconsistent estimates due to the so-called *incidental parameter problem.* However, in the PSM the focus is not on the consistency of the estimated coefficients of the propensity score model but on the balance it allows to achieve and in the consequent estimated ATT. Therefore, even if the dummy model suffers from the incidental parameter problem it could be appropriate for the estimation of the propensity score.

An alternative way to control for cluster effects in models for binary data is using *conditional logistic regression models*, that eliminates the cluster-specific effect by constructing a likelihood that is conditional on the number of treated in the cluster (Agresti, 2002). This approach solve the inconsistency of the dummy model but it is less efficient then the random effect model, especially when there are clusters containing only treated or controls; these clusters, in fact, are dropped from the analysis. Again, the inefficiency of the conditional logistic model should not affect the PSM performance. More problematic is the fact that since intercepts are not estimated using a conditional logistic regression, we could use this model to construct propensity score based distance measures only within clusters (Aussems, 2008). Therefore, we will not consider this method since we focus only on approaches which do not force a within-cluster matching.

Arpino and Mealli (2008) with a detailed Monte Carlo simulation experiment show that multilevel and dummy models help in reducing the biasing effect of unobserved cluster-level covariates and in particular dummy model show the best performance under several experimental conditions.

3.2.3 A weaker version of the SUTVA

There are several reasons that can make the SUTVA assumption untenable in a multilevel setting and these 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 results in section 4.2.

If we do not use the SUTVA, 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:

Assumption A.4. (A weaker version of SUTVA)

$$Y_{ij}(\boldsymbol{D}) = Y_{ij}(D_{ij},L_j);$$

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:

$$ATE^{D}_{|L=1} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid L=1) \tag{3}$$

$$ATE^{D}_{|L=0} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid L=0) \tag{4}$$

$$ATT^{D}_{|L=1} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid D=1, L=1) \tag{5}$$

$$ATT^{D}_{|L=0} = E\,(Y_{D=1}\text{-}Y_{D=0} \mid D=1, L=0) \tag{6}$$

The parameters (3) and (4) 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 (5) and (6) are the correspondent versions of parameters (3) and (4) 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:

$$ATE^{D} = ATE^{D}_{|L=1} * P(L=1) + ATE^{D}_{|L=0} * P(L=0) \tag{7}$$

$$ATT^{D} = ATT^{D}_{|L=1} * P(L=1 \mid D=1) + ATT^{D}_{|L=0} * P(L=0 \mid D=1) \tag{8}$$

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 called to check if the effect of the 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 (8) and (9) 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

section 3.1. 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 usually made in multilevel research about the need to disentangle within and between effects in multilevel models (Neuhaus and Kalbfleisch, 1998).

Parameters (3)-(8) 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 (7) and (8) are:

$$ATE^{L} = ATE^{L}_{|D=1} * P(D=1) + ATE^{L}_{|D=0} * P(D=0) \quad (9)$$

$$ATT^{L} = ATT^{L}_{|D=1} * P(D=1 \mid L=1) + ATT^{L}_{|D=0} * P(D=0 \mid L=1) \quad (10)$$

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:

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

$$0 < P(D=1|X, C, L) < 1 \; ; \; 0 < P(L=1|X, C, D) < 1. \quad (12)$$

Assumptions (11) and (12) are, basically, a generalization of the standard assumptions we have introduced in section 3.1 for the case of a single treatment. If we are not interested in all the parameters (3)-(10) we can use weaker versions of assumptions (11) and (12). 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 $0 < P(D=1|X, C, L=1) < 1$.

In order to estimate causal parameters (3)-(10) we can use a PSM procedure. For the estimation of parameters (3)-(8) we first need 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) and (5). The latter, will be employed for the estimation of parameters (4) or (6). Parameters (7) and (8) 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 propensity score specification discussed in the previous section.

4. Results

In this section we present the results of the estimation of the causal effect of childbearing on poverty by using propensity score matching methods. The conventional approximation for the household's welfare in Less Developed Countries is to use the household's observed consumption expenditure, which requires detailed information on consumption behaviour and its expenditure pattern (Coudouel et al. 2002; Deaton and Zaidi 2002). The expenditure variables are calculated by the World Bank procedure which is readily available with the VLSMS. We choose a relatively simple equivalence scale giving to each child aged 0-14 in the household a weight of 0.65 relative to adults[3].

Our sample is restricted to households where in the first wave consisted of at least one married woman aged between 15 and 40 years[4]. The selection is important since it avoids units who are in effect incapable of childbearing. Matching is based on the nearest neighbor method with replacement using the *nnmatch* module in *STATA* (Abadie et al. 2004)[5].

Before showing the estimates, we present a simple descriptive statistics in Table 1, demonstrating a clear negative *association* between number of children and consumption expenditures. We also see that if households experiencing at least one childbearing event have a lower average consumption growth than those without new children: 1004 dongs versus 1446 dongs. This simple difference would be a unbiased estimate of the causal effect of childbearing only if childbearing events were randomly assigned to families.

<Table 1 about here>

With the propensity score matching procedure we can control for a series of confounders. Our choice of variables is based mainly on dimensions which are important for both household's standard of living and fertility behaviour and hence are potentially confounders that have to be included in the conditioning set X to make the UNC plausible. All these variables can theoretically have an impact on change in consumption expenditure and on the decision to have children. Many

[3] We assessed the robustness of results to the imposed equivalence scale. Results are consistent to those presented here for reasonable equivalence scales. This analysis is available from authors upon request.

[4] This sample selection criterion is part of the matching strategy since we avoid comparing households having a child with households who were essentially out of the risk set (here because there are no women of fecund age in the household). Obviously different selection strategies are possible. However, this selection criterion gives low attrition with respect to households having additional children. Moreover, we tried the following alternative selection criterion: 1) select households with at least one married woman aged 15-35 in the first wave; 2) select households where the head or its spouse is a married woman aged 15-40 in the first wave; 3) select households where the head or its spouse is a married woman aged 15-35 in the first wave. However, results are very similar to those presented here.

[5] This software implements the estimators suggested by Abadie and Imbens (2002) and enables us to obtain analytical standard errors which are robust to potential heteroschedasticity. We prefer analytical standard errors to bootstrapped ones since Abadie and Imbens (2004) show that bootstrap fails with nearest neighbor matching.

of these variables are defined in terms of household ratios. That is, we include the number of household members that are engaged in gainful employment as a ratio of the total number of household members. We also include demographic characteristics of the household such as the sex and the age of the household head, the household size and the presence of existing children. The effect of children is further distinguished by their age distribution, and again expressed as a ratio of the total number of household members. Other covariates include 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, an education index, the level of equivalized consumption at the first wave and regional dummies. We also use two binary variables indicating, respectively, if the household is mainly engaged in farming or not and if the household head belongs to the majority ethnic group (the Kinh) or not. Importantly, we include also community information through three indexes: 1) an index of economic development, 2) health facilities and 3) educational infrastructures.

4.1. Causal effect of fertility on poverty comparing different strategies for the propensity score model specification

We start presenting results of the ATE and ATT estimates obtained using different propensity score models under the standard version of the SUTVA. Table 2 describes the different models we use for the propensity score estimation. 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.

<Table 2 about here>

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 inclusion of random slope could help to better balance some important covariates. The last models (9) is a single level logit models with dummy indicators for clusters.

We are not primarily interested in the goodness of fit[6] of these models, but in the balancing they allow us to achieve. As a measure of the balancing of the covariates we adopt the absolute

[6] 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

standardised bias (ASB) defined as the difference between the means of a covariate in the treated and control group divided by the average standard deviation in the two groups.

We want to stress two types of comparisons among these models. The first comparison is among single-level versus two-level logit and dummy models for the propensity score. 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 potentially unobserved community characteristics. This is the same advantage of model 9, where all the cluster-level heterogeneity (observed and not) is captured trough the dummies.

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 shown in the table 3 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 2, 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. 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.

<Table 3 about here>

At this point we consider the performance of the dummy model. As we can see from table 3, model 9 allows us to maintain a good balancing, on average, in both first and second level covariates. This strategy is the one to be preferred, because it allows to achieve the best balance in observed covariates and take also into account potentially important unobserved community variables. The superiority of dummy models in the specification of the propensity score for multilevel data is confirmed in a detailed Monte Carlo study by Arpino and Mealli (2008).

The second comparison we want to stress is among models that include and models that ignore community characteristics. 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 and to assess the importance of community information on the estimate. We get an interesting result. Models 4 and 8 allow to

---

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.

achieve a reasonably good balancing of community variables, even if these are not included in the matching set. On the contrary, the single level propensity score model that includes only first level covariates (model 2), since it ignore the clusterisation of households, does not achieve an acceptable balance of second level covariates. Finally, we already noted that the model with cluster indicators, that cannot include cluster-level variables, allow to obtain a good balancing in macro level characteristics. Concluding, model 9 seems to "dominate" all the others and is chosen as the "best" specification for the propensity score.

As the estimates are concerned, we can see from table 3, that ATT and the ATE obtained through a PSM procedure are not noticeably sensitive to the specification of the propensity score. Nevertheless, estimates based on the propensity score 9, which we choose as a reference, are a bit higher, in absolute value, than those obtained using a single-level propensity score model, that is the standard approach in applied works. This result is qualitatively in line with those obtained by the multilevel specifications, indicating that controlling for potentially unobserved community level confounders we get a slightly stronger estimated effect.

4.2. 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 3.2. 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 noticed in section 3.2, 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 9 of table 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 4, the estimated ATE and ATT for treatment *D* under the two versions of the SUTVA are quite similar. This is an indication that, the within and between effects are not significantly different, as also demonstrated by multilevel model estimates not reported here but available from the author upon request.

<Table 4 about here>

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 $ATT^{D}_{|L=0}$. Implications are discussed in the next section.

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[7]. Therefore, it seems that living in a community with a "high" level of

[7] 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*.

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).

5. Discussion and concluding remarks

Causal inference methods have been receiving increasing interest in the demographic literature. In this paper we stress the importance to keep explicitly into account the multilevel data structure when aiming at estimating causal effects in demographic studies when the clusterisation of units have substantial effect on the studied phenomena.

We illustrated the statistical and substantial issues related to causal inference in multilevel setting with reference to the estimation of fertility effects on poverty and focussing on the role of the community context. We show that ignoring relevant community information can have an important biasing effect on estimates.

We propose the use of multilevel specification and models with clusters indicators for the propensity score estimation as a mean to keep into account potentially important macro level unobserved confounders. We find that these methods allow to achieve a good balance in community level variables, but a simple model with dummies has to be preferred, at least in our application, since it allow us to maintain also a good balance of household level variables. An important result, useful in those situations where no cluster-level information is available, is that multilevel specification and the dummy model allowed to balance community-level variables even when these were not included in the matching set for the estimation of the propensity score model. Therefore, with this paper we suggest that practitioners should carefully consider the potential importance of unobserved contextual variables and give some guidelines on how to exploit the hierarchical structure of data to reduce potentially important biasing effects. We show that a simple model with cluster indicators, despite statistical prejudice one could have against it, serve quite well the scope of reducing potential bias due to omitted cluster level confounders.

We also discuss the potential violation of the SUTVA assumption in a multilevel setting and propose a simple method to allow some form of interference among units belonging to the same cluster.

From a substantial point of view, we tried to contribute to the literature on the relationship between fertility and poverty, which is a long contested issue among demographers and economists, by using a proper causal inference approach and adequate panel data on the rural Vietnam. We find a statistical significant and substantial causal effect of childbearing events in decreasing household consumption expenditures growth. The effect is robust to the specification of the propensity score and to the version of the SUTVA we adopt.

Interestingly, 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. On the other hand, we found no significative effect of living in a high-fertility community versus a low-level one. 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 this 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 higher.

Finally, we find that living in a high-level fertility community does not have a statistically significant effect of on poverty *per se*. In other words, we found that at the micro (household) level fertility has a negative causal effect on wellbeing, while at the aggregate level (community) there is a mere (spurious) association.

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., Kedir, A. M., Tadesse Weldegebriel, H. (2006) "State Dependence and Causal Feedback of Poverty and Fertility in Ethiopia", ISER Working Paper No 2006-30.

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.

Admassie, A. (2002) Explaining the High Incidence of Child Labour in Sub-Saharan Africa. *African Development Review,* 14(2): 251 – 275.

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*., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association,* 91, 444–472.

Anh, *D*. N. and Thang, N. M. (2002) Accessibility and Use of Contraceptives in Vietnam. *International Family Planning Perspectives*, 28(4), 214-219.

Arpino (2008) Causal Inference for Observational Studies Extended to Multilevel Settings. The Case of the Fertility Effect on Poverty in Vietnam. Unpublished PhD thesis, Department of Statistics, University of Florence.

Arpino and Aassve (2008) Estimating the causal effect of fertility on economic wellbeing: Data requirements, identifying assumptions and estimation methods. Dondena Working Paper n.13.

Arpino and Mealli (2008) The specification of the propensity score in multilevel observational studies. Dondena Working Paper n.6.

Aussems, C. (2008) Multilevel data and propensity scores: An application to a Virtual Y after-school program, mimeo.

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.

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.

Caliendo, M. and Kopeining, S. (2005) *Some Practical Guidance for the Implementation of Propensity Score Matching*. IZA working paper, 1588.

Coudouel, A., Hentschel, J. and Wodon, Q. (2002) *Poverty Measurement and Analysis*, Poverty Reduction Strategy Paper Sourcebook, World Bank, Washington *D*.C.

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.

Dawid, A. P. (1979) Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society B*, 41, 1-31.

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.

Drovandi, S. and Salvini. S. (2004) Women's Autonomy and Demographic Behavior. *Population Review,* 43(2).

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. and Crimmins E.M. (1985) *The Fertility Revolution.* Chicago: University of Chicago Press.

Engelhardt, H., Kohler, H-P. And Fürnkranz-Prskawetz, A. (Eds.) (2009), *Causal Analysis in Population Studies. Concepts, Methods, Applications.* Springer.

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.
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.
Fisher, R. A. (1925) *Statistical Methods for Research Workers*. 1st Edition. Oliver and Boyd, Edinburgh.

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.

Goodman, A. and Sianesi, B. (2005) Early Education and Children's Outcomes: How long the impacts last. *Fiscal Studies*, 26(4).

Goldstein, H. (1995) *Multilevel Statistical Models*, Edward Arnold, London.

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.

Haughton, D., Haughton, J. and Phong, N. (edited by) (2001) *Living Standards during an Economic Boom. Vietnam 1993-1998*, Statistical Publishing House, Hanoi.

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. (1997) Instrumental Variables: A study of implicit behavioural assumptions used in making program evaluations. *Journal of Human Resources*, 32, 441-462.

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.

Hirschman, C. and Guest, P. (1990) Multilevel models of fertility determination in four Southeast Asian countries: 1979 and 1980. *Demography*, 27 (3), 369-396

Imbens, G. W. (2000) The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika,* 87(3), 706-710.

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.

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 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.

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.

Li, F., Zaslavsky, A.M. and Landrum, M. B. (2009) Propensity score analysis with hierarchical data, mimeo.

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.

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.

Mukherjee, S. and Benson, T. (2003) The Determinants of Poverty in Malawi, 1998. *World Development*, 31(2), 339 – 358.

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).

Neuhaus, J.M. & Kalbfleisch, J.D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54, 638-645.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

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.

Skinner, B. F. (1965) *Science and human behaviour*, New York: Free Press.

Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*, London: Sage.

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.

Stuart, E. A. (2007 Estimating causal effects using school-level datasets. *Educational Researcher*, 36, 187-198.

Su, Y. (2008) Causal Inference of Repeated Observations: A Synthesis of Matching Method and Multilevel Modeling. Paper presented at the annual meeting of the APSA 2008, Hynes Convention Center, Boston, Massachusetts.

Van de Walle, D. (1996) Infrastructure and Poverty in Vietnam, Living Standards Measurement Study Working Paper No. 121, The World Bank Group, Washington DC.

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.

Table 1. Average equivalized household consumption expenditure at the two waves and its growth by number of children born between the two waves.

Number of children born between the two waves	Observations	Average consumption in 1992	Average consumption in 1997	Average consumption growth in 1997-1992
0	1232	970	2436	1466
1	581	856	1892	1036
2	182	790	1755	965
3	28	571	1154	583
At least 1	791	832	1835	1004
Total	2023	916	2201	1285

Note: We consider the number of children of all household members born between the two waves and still alive at the second wave. All consumption measures are valued in dongs and rescaled using prices in 1992. The 2023 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.

Table 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 of household members)
8	With X, without C; RI; RS for *edu*
III - Single level logit model with clusters indicators	
9	With X and dummy indicators for clusters (“dummy model”)

Note: PS = propensity score specification; X = household level covariates; C = community level covariates; RI = random intercept; RS = random slope

Table 3. Comparison among different propensity score specifications: balancing and estimates.

Propensity score	Absolute Standardised Bias after matching				Estimates (thousands of dongs)	
	Household level covariates		Community level covariates			
	Mean	Median	Mean	Median	ATE	ATT
I - Single level logit models						
1	3.7	2.9	4.5	4.7	-411	-356
2	5.4	4.4	10.4	9.2	-421	-351
II – Two-level logit models						
3	6.0	4.8	3.6	2.8	-492	-431
4	7.0	6.2	3.4	3.2	-541	-470
5	5.5	3.5	1.5	1.2	-434	-384
6	5.7	3.9	4.2	3.8	-464	-514
7	5.7	5.7	4.4	4.7	-450	-375
8	6.3	5.3	6.8	3.8	-447	-407
III - Single level logit model with clusters indicators						
9	3.8	3.2	3.7	4	-458	-397

Note: all the estimates (ATE and ATT) are signifant at the 5% level.

Table 4. 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}	-458 (95)	ATT^{D}	-397 (91)
Estimates obtained under the weaker version of the SUTVA			
$ATE^{D}_{\mid L=1}$	-420 (146)	$ATT^{D}_{\mid L=1}$	-566 (176)
$ATE^{D}_{\mid L=0}$	-447 (113)	$ATT^{D}_{\mid L=0}$	-313 (107)
ATE^{D}	-440 (90)	ATT^{D}	-425 (100)

Note: $L = 1$ for high-level fertility communities; $L = 0$ for low-level fertility communities. The estimates are all obtained using a PSM procedure following the specification used in model 9 (see table 2). The matching method employed is the nearest neighbour.