

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

Bruno Arpino

Abstract The thesis addresses some statistical issues arising in studies of causal inference with observational hierarchical data with an application on the estimation of the effect of fertility on poverty using Vietnamese data. In the thesis, I first contrast matching and instrumental variables methods stressing the fact that their validity relies on the plausibility of the underlying assumptions; which in turn depend on the available data. In the application, the multilevel dimension introduces statistical complications and stimulates interesting research questions. From a methodological point of view, I identify and discuss some issues that are not tied to the considered application but have a general relevance: the role of clusters in the assignment mechanism and the potential violation of SUTVA. The thesis proposes combinations of matching methods and multilevel modelling techniques and a weaker version of the SUTVA assumption. I find a negative and substantial effect of fertility on wellbeing; this effect is stronger in high-level fertility communities.

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

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.

This is an extended abstract of the author PhD thesis written at the Department of Statistics, Università di Firenze. The author would like to thank his supervisors Fabrizia Mealli, Arnstein Aassve and Letizia Mencarini.

1 Introduction

In many field of social research investigators are interested in drawing causal conclusions about the relationship existing among the variables under study. The potential outcomes framework for causal inference, which was pioneered by Neyman (1923) and extended by Rubin (1974) to observational studies has become increasingly popular. In particular, propensity score matching methods have been widely adopted.

Many applications use data with a multilevel structure but this aspect is often not considered explicitly. 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., Snijders and Bosker, 1999). Multilevel models have also been used for the estimation of causal effects. For example, Deheija (2003) discusses the use of hierarchical models for the evaluation of programs implemented at multiple sites. However, the study of the specific statistical issues for causal inference arising in multilevel settings is an underdeveloped field of research. The main thesis goal is to discuss in a unified and systematic way the main statistical complications implied by the multilevel data structure and give some simple advices to practitioners to deal with them. In particular, the thesis proposes combinations of matching methods and multilevel modelling techniques and a weaker version of the traditional SUTVA assumption. Hierarchical data structures are considered but the extension to non-nested structures is possible with minor modifications of the proposed approaches

The methodological issues are discussed and applied to the problem of estimating the causal effect of fertility on poverty in Vietnam. As it is often the case, in this application the multilevel dimension, represented by the clusterisation of households in communities, not only introduces statistical complications but also stimulates interesting research questions.

2 Causal inference under unconfoundedness in multilevel settings

I slightly modify the standard setup of the Rubin Causal Model to accommodate explicitly hierarchical data structures. Let consider a 2-level population of N individual units under study $ij = 1, 2, \dots, n_j$ grouped into clusters $j = 1, 2, \dots, J$. Let D represents a binary treatment, which assumes the value 1 for treated units and 0 for untreated units, while Y indicates the outcome. Under the Stable Unit Treatment Value Assumption (SUTVA, Rubin 1980) each unit, ij , has two potential outcomes: Y_{ij1} if $D_{ij} = 1$ and Y_{ij0} if $D_{ij} = 0$. This assumption requires that potential outcomes for a unit are not affected by the treatment received by other units and there are no versions of the treatment. In the following, I discuss two issues arising in causal inference with multilevel data: 1) The multilevel nature of the selection process and 2) Potential violation of the SUTVA.

2.1 The multilevel nature of the assignment mechanism

A fundamental ingredient in the Rubin Causal Model is the assignment mechanism. This is defined as the conditional probability of receiving the treatment, as a function of potential outcomes and observed covariates. First of all, having a hierarchical structure of the data would imply considering if the assignment mechanism works at the cluster or at the individual level. In the following, I will assume that the treatment is assigned at the individual level but cluster level characteristics enter in the assignment mechanism. In this thesis, I focus on observational studies where the assignment mechanism is unconfounded, meaning that conditional on a set of observed covariates the (unknown) assignment probabilities do not depend on the potential outcomes. In these settings, propensity score methods (Rosenbaum and Rubin 1983) that can be used to compare treated and control units with similar characteristics. The propensity score is defined as the conditional probability of receiving a treatment given pre-treatment characteristics and is usually estimated using a logit or probit model.

Rubin (2008) argues that “a crucial idea when trying to estimate causal effects from an observational dataset is to conceptualize the observational dataset as having arisen from a complex randomized experiment, where the rules used to assign the treatment conditions have been lost and must be reconstructed”. A convenient experimental design to be considered for 2-level data is the split plot design: in the first stage each of the J clusters is assigned to a treatment assignment strategy (that can be conceived as the percentage of units to be treated in the cluster); in the second stage within each cluster units are assigned to treatment or control. Having this hypothetical randomized experiment in mind, it turns out natural to introduce explicitly cluster characteristics when we write the identifying assumptions:

$$\begin{aligned} \text{Unconfoundedness:} \quad & Y_1, Y_0 \perp D \mid X, C \\ \text{Common support:} \quad & 0 < P(D=1 \mid X, C) < 1, \end{aligned}$$

where X and C represent, respectively individual and cluster level confounders.

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. 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. Within-cluster matching guarantees a perfect balancing in all the observed and unobserved cluster characteristics but is likely to be unfeasible in those situations, representing the norm in large scale observational studies, where a relatively few units within each cluster are present. Here I propose to take into account the multilevel data structure by using multilevel or fixed effects models. I compare these multilevel specifications with single level ones.

2.2 A weaker version of the SUTVA

The assumption of no interference between units (one of the two components of the SUTVA) is often made without critical examination. Recently the topic has received some attention in the causal inference literature (e.g., Hudgens and Halloran 2008).

In multilevel research problems the SUTVA assumption is often untenable. However, the structure of the data itself can suggest simple ways to overcome the problem. The consideration that SUTVA is intrinsically untenable with multilevel data is strictly related to the nature itself of the multilevel research that recognises a strong influence of the context on individuals. The notion of context is quite general and can include spatial contexts (e.g. countries or communities), organizational contexts (classrooms, schools, firms). In many applications is natural to think that the behaviour of one unit can produce some effect on the outcome of others units in the same cluster. However, it is also often reasonable to limit the interference within clusters and assume no interference among units belonging to different clusters. For example, it turns out natural to think that interventions given to certain students may affect other students results in the same class or that demographic behaviours of households can influence on other households economic outcomes in the same neighbourhood.

In general, without SUTVA potential outcomes for each units ij depend on the entire $N \times 1$ vector of treatments indicator, $\mathbf{D}$. 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: $Y_{ij}(\mathbf{D}) \equiv Y_{ij}(D_{ij}, \mathbf{D}_{-ij})$; where $\mathbf{D}_{-ij}$ 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. 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 : $Y_{ij}(\mathbf{D}) = Y_{ij}(D_{ij}, \mathbf{D}_{ij}^{(j)})$; where $\mathbf{D}_{ij}^{(j)}$ represents the treatments received by all units in the cluster j with the unit ij removed. 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 . Even if the number of potential outcomes is reduced it remains intractable in general settings. For example, in a clusters of size equal to 10 the potential outcomes are $2^{10} = 1024$ and they fast increase with the cluster size. To make the problem tractable it can be assumed that the relevant information contained in the vector $\mathbf{D}_{ij}^{(j)}$ can be summarized by a scalar function representing the proportion of treated in the cluster (calculated excluding the unit ij), $\mathbf{P}_{ij}^{(j)}$. Therefore, potential outcomes for unit ij can be written as a function of the treatment the unit receives and the proportion of the other units treated in the cluster: $Y_{ij}(\mathbf{D}) = Y_{ij}(D_{ij}, \mathbf{P}_{ij}^{(j)})$. This solution, similar to the approach followed by Hong and Raudenbush (2006), reduces the argument of potential outcomes to a 2-dimensional space. To further simplify the analysis, we can split the range of $\mathbf{P}_{ij}^{(j)}$ in a limited number of intervals. The information contained in $\mathbf{P}_{ij}^{(j)}$ will be summarised by a function, f , taking s values:

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

where $t_1, t_2, \dots, t_{s-1}$ are real numbers satisfying: $0 < t_1 < t_2 < \dots < t_{s-1} < 1$. In this way, potential outcomes are reduced to 2^s . The most simplified analysis, with the minimum number of potential outcomes, fixes s 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: $Y_{ij}(\mathbf{D}) = Y_{ij}(D_{ij}, L_j)$. Hence, we have only 4 potential outcomes according to the treatment the unit receives and to the level of proportion of treated in the cluster. Each contrast between two of the 4 potential outcomes define a causal parameter of possible interest. For example, it could be of interest to estimate the average effect of the treatment for units in clusters with a “high” percentage of treated, $ATE^{L=1} = E(Y_{11} - Y_{01})$ and compare this result with the average effect in clusters with a “low” presence of treated, $ATE^{L=0} = E(Y_{10} - Y_{00})$.

3 Data

To illustrate the statistical issues discussed in section 2, I use data from a 2-wave panel survey, the Vietnamese Living Standard Measurement Survey (VLSMS) conducted in 1992 and in 1997. The treatment variable is represented by the occurrence of at least one childbearing event between the two waves and the outcome is a measure of economic wellbeing based on the equivalised household consumption expenditures growth between the two waves, measured in 1992 Vietnamese dong. The VLSMS includes 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. Having data on two waves allows for a combination of matching and difference-in-difference techniques. In fact, the outcome variable can be defined as the difference between the equivalised household consumption expenditures after and before the treatment.

A very interesting feature of the VLSMS is that it also provides detailed community information from a separate data file. Community level information is available for rural areas only and includes 120 communities, with information on markets, roads, hospitals and other important infrastructures and main economic activities. Our sample is restricted to households living in the rural areas with at least one married woman aged between 15 and 40 years in the first wave. This selection is part of the matching strategy since it avoids units who are in effect incapable of childbearing.

4 Results and discussion

Given the space constraint I will only show the results of the analysis where the multilevel dimension is taken into account. Results from instrumental variable techniques and standard propensity score matching are available from the author upon request. Data have a 2-level structure with household at the first level and communities at the second. The communities sample sizes ranges from 7 to 21 making unfeasible to use within-cluster matching procedures. The communities effects will be kept into account by using random and fixed effect models for the propensity score. The different specifications I compare are described in Table 1. The aim of these models is to achieve a good balancing in the covariates between the groups of treated and control units. As a

measure of the balancing I use the absolute standardized bias (ASB) after matching: the lower the ASB the better the balancing.

As can be seen from Table 2, single level logit models, where the multilevel dimension is not taken into account, do not guarantee an acceptable balancing of community level confounders. The multilevel propensity score models show an improvement in the balancing of community level variables but individual level variables are often not well balanced. Finally, a simple logistic regression with dummy indicators for communities, allows to obtain a good balancing of both first and second level covariates. Therefore, this strategy is the one to be preferred because it allows to balance observed covariates and potentially important unobserved community factors.

Table 1: Description of the propensity score specifications used in the analyses.

Description
<i>I - Single level logit models</i>
1 With X and C
2 With X, without C
<i>II - Two-level logit models</i>
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
<i>III - Single level logit model with clusters indicators</i>
9 With X and dummy indicators for clusters (“dummy model”)
Note: X = household level covariates; C = community level covariates; RI = random intercept; RS = random slope.

This is confirmed by a Monte Carlo simulation study, where I find that logistic regressions with dummies reduce the bias on the Average Treatment Effect due to omitting important cluster level confounders. With the specification of the propensity score matching I choose on the basis of the previous considerations (see n. 9 of Table 2), it is estimated that, on average, a childbearing event decreases household consumption expenditures growth by 397 thousands dong. This effect is not only statistically significant but also economically substantial if we consider, for example, that the average consumption growth between the two waves was equal to 1285 thousand dong.

Finally, I compare the estimates obtained under the standard version of the SUTVA, used in the previous analysis, with those I obtain allowing for within-community interference. The most interesting result shown in Table 3 is that the average causal effect of fertility on poverty for treated households living in high-fertility communities, is higher, in absolute value, than the same quantity estimated in low-fertility communities. This result is likely to be due to the within-community competition for and share of resources. More explicitly, the facilities, services (e.g. provided by health care or family planning centers), benefits (e.g. maternity benefits),

which are made available in a community to help households with children are subject to economic constraints. In communities where the number of children is higher these limitations generate competition among households. It is more 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 areas.

Table 2: Comparison among different propensity score specifications: balancing and estimates.

Propensity score specification	Absolute Standardised Bias after matching				Estimates (1000 of dong)	
	Household covariates		Community covariates		ATE	ATT
	Mean	Median	Mean	Median		
<i>I - Single level logit models</i>						
1	3.7	2.9	4.5	4.7	-411	-356
2	5.4	4.4	10.4	9.2	-421	-351
<i>II – Two-level logit models</i>						
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
<i>III - Single level logit model with clusters indicators</i>						
9	3.8	3.2	3.7	4.0	-458	-397
Note: all the estimates (ATE and ATT) are significant at the 5% level.						

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. As far as the effect of the community fertility level per se, its average causal effect is equal to -101 with a standard error of 162. Therefore, it seems that living in a community with a “high” level of fertility or in a community with a “low” level of fertility does not affect the households’ living standard significantly. This, obviously, after controlled for compositional and contextual differences existing among communities, captured by the control variables. Therefore, the negative association it is often found 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, I 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. I also plan in the near future to extend the analysis accordingly to what has been discussed in section 2.3.

Table 3: Estimated causal effects of childbearing events on poverty under two versions of the SUTVA.

	ATE		ATT
<i>Estimates obtained under the standard version of the SUTVA</i>			
ATE ^D	-458 (95)	ATT ^D	-397 (91)
<i>Estimates obtained under the weaker version of the SUTVA</i>			
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)

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 1). The matching method employed is the nearest neighbour.

References

1. Dehejia, R.: Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data, *Journal of Business & Economic Statistics*, 21(1), (2003).
2. Hong, G., Raudenbush, S.W.: Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *JASA*, 101(475), 901-910, (2006).
3. Hudgens, M.G., Halloran, E.M., Toward Causal Inference With Interference, *JASA*, 103, (482), (2008).
4. Kim, J., Seltzer, M.: Causal Inference in Multilevel Settings in which Selection Process Vary across Schools. Working Paper 708, Center for the Study of Evaluation (CSE): Los Angeles, (2007).
5. Neyman, J.: On the application of probability theory to agricultural experiments: essay on principles, section 9, (1923). Translated in *Statistical Science*, 5(4), 465-480, (1990).
6. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55, (1983).
7. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701, (1974).
8. Rubin, D.: Discussion of Randomization Analysis of Experimental Data: The Fisher Randomization Test by D.Basu. *JASA*, 75, 591-93, (1980).
9. Rubin, D.B.: For Objective Causal Inference, Design Trumps Analysis, *The Annals of Applied Statistics*, 2 (3), 808-840, (2008).
10. Snijders, T.A.B., Bosker, R.J.: *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*, London: Sage, (1999).