The specification of the propensity score in multilevel observational studies

Arpino, Bruno and Mealli, Fabrizia

Università Commerciale "Luigi Bocconi" - Department of Decision Sciences, Università Commerciale "Luigi Bocconi" - DONDENA research centre, Università di Firenze - Dipartimento di Statistica

2008

Carlo F. Dondena Centre for Research on Social Dynamics

DONDENA WORKING PAPERS

The specification of the propensity score in multilevel observational studies

Bruno Arpino, Fabrizia Mealli

Working Paper No. 6
URL: www.dondena.unibocconi.it/wp6

October 2008

Carlo F. Dondena Centre for Research on Social Dynamics
Università Bocconi, via Guglielmo Röntgen 1, 20136 Milan, Italy
http://www.dondena.unibocconi.it

The opinions expressed in this working paper are those of the author and not those of the Dondena Centre which does not take an institutional policy position.

ISSN 2035-2034

The specification of the propensity score in multilevel observational studies

Bruno Arpino
Carlo F. Dondena Centre for Research on Social Dynamics
Università Bocconi
via Guglielmo Röntgen 1
20136 Milan
Italy
bruno.arpino@unibocconi.it

Fabrizia Mealli
Department of Statistics
University of Florence
Viale Morgagni 59
50134 Florence
Italy
mealli@ds.unifi.it

Abstract

Propensity Score Matching (PSM) has become a popular approach to estimation of causal effects. It relies on the assumption that selection into a treatment can be explained purely in terms of observable characteristics (the "unconfoundedness assumption") and on the property that balancing on the propensity score is equivalent to balancing on the observed covariates. Several applications in social sciences are characterized by a hierarchical structure of data: units at the first level (e.g., individuals) clustered into groups (e.g., provinces). In this paper we explore the use of multilevel models for the estimation of the propensity score for such hierarchical data when one or more relevant cluster-level variables is unobserved. We compare this approach with alternative ones, like a single level model with cluster dummies. By using Monte Carlo evidence we show that multilevel specifications usually achieve reasonably good balancing in cluster level unobserved covariates and consequently reduce the omitted variable bias. This is also the case for the dummy model.

Keywords

propensity score, multilevel studies, unconfoundedness, causal inference

1. Introduction

In many fields of the social sciences, there is a growing interest in methods that can be used to evaluate the effects of social programs and public policies. A large part of the recent literature on program evaluation focuses on estimation of the average effect of the treatment under the potential outcomes framework for causal inference, which was pioneered by Neyman (1923) and Fisher (1925) and extended by Rubin (1974, 1978) to observational studies. Following the seminal work of Rosenbaum and Rubin (1983a) the literature on estimating average treatment effects under the unconfoundedness assumption has become widespread (Imbens, 2004). In particular, propensity score matching methods have become popular among researchers: this approach is, for example, widely applied when evaluating labour market policies (see e.g., Bryson et al, 1992; Heckman et al, 1997; Dehejia and Wahba, 1999; Manski and Garfinkel, 2002; Sianesi, 2004). In the empirical labour economics literature, matching has been used to evaluate the returns from education (e.g., Blundell et al, 2005; Brand and Halaby, 2006) and the union membership wage premium (e.g. Bryson, 2002; Eren, 2007). Empirical examples can be found in very diverse fields of observational studies whenever the researcher aims to evaluate the effect of a variable (often of some policy relevance) on another. In the demo-economic literature, researchers are often interested in the evaluation of the effects of demographic events, like childbearing or marital disruption, on economic variables, like wellbeing and labour force participation (e.g., Aassve et al, 2007; Aassve and Arpino, 2007). The approach is also applied in the educational literature to study the effect of educational programs and policies on students' performances (e.g., Hong and Raudenbush, 2006).

In all these applications, the possible occurrence of selection bias needs to be discussed and addressed. In fact, taking the mean outcome of non-participants as an approximation for absence of treatment is not advisable, since participants and non-participants usually differ even in the absence of treatment. This is the well-known *selection bias* problem and one possible solution is the matching approach: the basic idea is to find in the group of non-treated units, those individuals who are as similar as possible to the treated subjects in all relevant pre-treatment characteristics, *X*. Once this is done, differences in outcomes for this purposefully selected control group and of participants can be attributed to the treatment. The underlying assumption is known as *unconfoundedness.* In practice, we attempt to identify a set of pre-treatment characteristics, *X*, which includes all observable variables affecting both the outcome and the treatment assignment. Since conditioning on all relevant covariates is difficult in the case of a high dimensional set *X* (*curse of dimensionality* problem), Rosenbaum and Rubin (1983a) suggest the use of a *balancing score*, *b*(*X*), a function of the observed covariates *X*, such that the conditional distribution of *X* given *b*(*X*) is independent of assignment to treatment. The *propensity score*, the probability of participating in a treatment given the observed characteristics, *X*, is a balancing score. The resulting matching technique is known as propensity score matching (PSM).

Obviously, PSM cannot solve the evaluation problem in every case. As noted by Blundell et al (2005), it should only be applied if the unconfoundedness assumption can be credibly invoked based on the informational richness of the data and a detailed understanding of the institutional set-up by which selection into treatment takes place. In fact, the underlying identification assumption, unconfoundedness, rules out the role of the unobservable variables. If any of the relevant covariates is unobserved, PSM estimates will be biased.

The issue of selection on unobservables, without moving to instrumental variable methods, has been addressed using models for sensitivity analysis (e.g., Rosenbaum and Rubin, 1983b; Ichino et al, 2008) or by means of non parametric bounds for treatment effects (Manski, 1990).

In this paper we address the selection bias problem due to unobserved covariates in a specific setting, that of multilevel studies where the unobserved covariates are at the higher (cluster) levels. Multilevel studies analyse multilevel structured populations, which are the norm in many fields. Education provides a prototype example. Pupils or students are grouped in classes; classes are nested within schools; and schools may be administered within local authorities or school boards. The units in such a system lie at four different levels of a hierarchy: pupils are assigned to level 1, classes to level 2, schools to level 3 and authorities or boards to level 4. Other examples of hierarchical populations are people within households, within geographical areas of residence (which are often found in the demographic and socio-economic research); workers within firms within local labour markets (which are the typical structure in the labour economic literature). Multilevel research analyses the interrelationship existing between the different levels and takes into account the variability associated with each level of the nesting. Multilevel modelling techniques have been used to bring together, simultaneously, macro and micro level variables while accounting for the dependence of observations within groups (see e.g., Goldstein, 1995; Hox, 1995; Snijders and Bosker, 1999).

In the paper we explore the use of multilevel techniques for the specification of the propensity score for multilevel data. The issue has received little focus in the literature. For the best of our knowledge only Kim and Seltzer (2007) address the issue explicitly. They propose use of a multilevel model for estimation of the propensity score and then implementation of the matching algorithm within each cluster. If we impose the condition that treated and matched controls must belong to the same cluster, we then automatically achieve perfect balancing in all the observed and unobserved cluster characteristics. This strategy is not likely to be feasible 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 belonging to the same cluster.

We distinguish between two assignment mechanisms, and the consequent version of the unconfoundedness assumption needed for the identification of causal effects, that differ according to the way the cluster effects enter the selection into treatment process. In the first case, the cluster characteristics along with individual ones affect the selection probability. In the second situation, the selection process differs by cluster and belonging to a cluster instead of another leads the individual probabilities to be selected to vary. This distinction is relevant not only conceptually but also for the practical implementation of the matching procedure. In fact, in the first case within-cluster matching is not needed since what we require is that matched treated and control units belong to *similar* clusters and not necessarily to the *same* cluster. In this paper we focus on this kind of setting and on typical large-scale surveys (e.g. national surveys often characterised by a two-stage sampling scheme), where the cluster sizes are generally small. For this setting, we explore the use of multilevel specifications for the propensity score without imposing a within-cluster matching requirement. We also propose a more direct way to control for unobserved heterogeneity at the cluster level: this consists in a two-stage procedure, where in the first stage we estimate a multilevel model for the selection process and obtain predictions of the random error at the cluster level then, at the second stage, we estimate a single level model for the propensity score including as an additional covariate the predictions of the random effects obtained at the previous stage. We will refer to this method as the *two-stage procedure*. Both the multilevel specification of the propensity score and the two-stage procedure should help to mitigate the biasing effect of unobserved macro level covariates. We compare these multilevel specifications with single level ones by using a series of Monte Carlo experiments, where the interest lies in the bias of the Average Treatment effect on the Treated (ATT) estimators. We also compare multilevel specifications with alternative approaches which try to keep into account for 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 (and omitting the overall intercept). We will refer to this approach as the *dummy* *model.* 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 PSM the focus is not on the consistency of the estimated coefficients of the propensity score model but on the balance it obtains 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 the *conditional logistic regression*, 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 resolves the inconsistency of the dummy model but is less efficient than the random effect model, especially when there are clusters containing only treated or controls; these clusters, in fact, cannot be considered in the analysis. In addition, 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. Therefore, we will not consider this method since we focus only on approaches not forcing a within-cluster matching.

The paper is organised as follows. Section 2 sets up the framework using the potential outcomes framework and provide a motivation for the paper by means of examples where the topic could be relevant. Section 3 discusses some different multilevel setting and alternative propensity score specifications. Section 4 provides Monte Carlo results for the performance of these specifications and section 5 concludes.

2. Framework and motivation

The propensity score matching methodology for the estimation of causal effects

We use a standard setup in the treatment effect literature. Let us suppose we have a population of individual units under study indexed by $i = 1, 2, \ldots , N$, an indicator for a binary treatment, T, which assumes the value 1 for treated units and 0 for untreated, or controls, and an outcome variable, which we indicate by Y. Under the potential outcomes framework, each unit, i, has two potential outcomes associated with the two treatment levels: Y_{i1} if $T_i = 1$ and Y_{i0} if $T_i = 0$. Potential outcomes for unit i and treatment t can be written as Y_{id}, with $t = \{0,1\}$. The fact that this variable is labelled only by i and t corresponds to the “no interference among units” assumption of Cox (1958), which Rubin (1980) extended to the Stable Unit Treatment Value Assumption (SUTVA). This standard 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. Under the SUTVA, we can define several causal effects the most popular being the Average Treatment effect on the Treated:

$$\text{ATT} = E(Y_1\text{-}Y_0|\ T = 1), \tag{1}$$

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 than the average treatment effect on the whole population (Heckman et al, 1997). The identifying assumptions are usually stated as follows:

Unconfoundedness (A.1): $Y_1, Y_0 \perp T|X$,
Common support (A.2): $0 < P\,(T=1|\,X) < 1$,

where ⊥ in the notation introduced by Dawid (1979) means independence. The combination of the two assumptions A.1 and A.2 is referred to as *strong ignorability* (Rosenbaum and Rubin, 1983a). Assumption A.1, known as the unconfoundedness assumption, asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates. In other words, within subpopulations defined by values of the covariates, we have random assignment. This assumption rules out the role of the unobservable variables and therefore is referred to also as selection on observables (Imbens, 2004). Assumption A.2, known as the common support assumption, implies equality in the support of X in the two groups of treated and controls (i.e. Support($X|T$=1) = Support($X|T$=0)) which guaranties that the ATT is well defined (Heckman et al, 1997); otherwise, for some values of the covariates there would be some treated for which we could not find any comparable units in the control group.

It is instructive to remember the decomposition of the selection bias proposed by Heckman et al (1998). They showed that the selection bias (B) can be decomposed in three components: $B = B_1 + B_2 + B_3$. The first component, B_1, refers to the bias caused by non-overlapping supports of X in the treated and control group. The term B_2 depends on misweighting within the common support, as the empirical distributions of treated and non-treated may not be the same even when restricted to the common support. Finally, the term B_3 is the "true econometric selection bias" resulting from "selection on unobservables", that is, it is the bias arising from a different distribution of relevant unobserved variables between treated and controls. Under A.1 the term B_3 is zero. The other bias components are cancelled out when we restrict the analysis on the common support (B_1) and we balance covariates in the group of treated and control units (B_2).

Several methods are available to balance covariates across the groups of treated and controls. Among them matching has become very popular. Matching is an intuitive and appealing method, which basic idea consists of contrasting treated and control units with the same characteristics X. Starting from assumption A.1, the basic idea is that within each cell defined by the values of the covariate X assignment to treatment or control group is random. Therefore, if in a given application we are willing to assume that all relevant variables that affect the selection on treatment and outcome are collected in the set X (and hence we are confident that assumption A.1 holds) we can match each treated unit with one (or more) control unit with the same values of X. The group of treated and matched controls will differ only for the exposure to treatment and, therefore, differences in the outcome between the two groups can be attributed to the treatment. When the number of matching variables is large and/or when some of X are continuous exact matching becomes unfeasible and a distance metric have to be used to weight comparisons of matched treated and control units. An alternative is to implement the matching on a univariate variable, which "summarizes" the information incorporated in X, as opposed to matching directly on the multivariate set X. Well known are matching methods that use the *propensity score*, which can be defined as the conditional probability of receiving a treatment given pre-treatment characteristics:

$$e(X) \equiv Pr\{T = 1|X\} = E\{T|X\}. \tag{2}$$

The substitution of the multivariate set X with the univariate $e(X)$ in the matching procedure is justified by two important theorems due to Rosenbaum and Rubin (1983a). The first one, referred to as the balancing property of the propensity score, asserts that conditioning on the propensity score, X and T are independent: $X \perp T \mid e(X)$. This result implies that observations with the same propensity score have the same distribution of characteristics X, independently of treatment status. When the propensity scores are balanced across the treatment and control groups, the distribution of all the covariates are balanced in expectation across the two groups. Therefore, matching on the propensity score is equivalent of matching on X. The second theorem shows that if treatment assignment is strongly ignorable given X, then it is strongly ignorable given any balancing score, then adjusting for $e(X)$ is sufficient to produce unbiased estimates of the ATT. On the basis of these two theorems we can write the ATT as:

$$\text{ATT} \;= E_{e(X)\mid T=1}\left[E\left(Y_1 \mid T=1, e(X)\right) - E\left(Y_0 \mid T=0, e(X)\right)\right] \tag{3}$$

where the outer expectation is over the distribution of *e(X)* in the sub-population of the treated units.

In observational studies the propensity score is not known and it has to be estimated from the data available. Using the common logit or probit models, we can write $e(X) \equiv Pr\{T = 1|X\} = F[h(X)]$, where $F(.)$ is, respectively, the normal or the logistic cumulative distribution and $h(X)$ is a function of covariates with linear and higher order terms. The choice of which higher order terms to include, as well as interactions among covariates, is determined solely by the need to balance covariates distribution in the two treatment groups (Dehejia and Wahba, 1999). Simple parametric specifications for the propensity score have indeed often been found to be quite effective in achieving the balancing required (see for example Zhao, 2005).

The estimation of the propensity score is, however, not sufficient to estimate ATT using the (3). The reason is that the probability of observing a treated and a control unit with exactly the same value of the propensity score is, in principle, zero, since $e(X)$ it is a continuous variable. Then, we need to use some algorithm to match treated and controls. Various matching methods have been proposed in the literature to overcome this problem and the most widely used are nearest neighbour, stratification, radius, kernel matching (see e.g., Caliendo and Kopeining, 2008).

Motivation: The “unmeasured context” problem

As said before, the PSM methodology is based on the validity of the unconfoundedness assumption. If one or more variables affecting the selection into treatment and potential outcomes are not observed, making the unconfoundedness assumption to fail, the estimated ATT will be biased.

In this paper we consider a particular case of omitted variable bias caused by one or more unobserved cluster level covariate in a multilevel data structure. We refer to this issue as the “unmeasured context” problem. To give a concrete exemplification we consider the case of the evaluation of labour market programs. Workers belonging to the same local labour market share the same institutional, cultural and socio-economic environment and, as a consequence, they are likely to show, *ceteris paribus* (with respect to individual characteristics), more similar probabilities to be selected into the program and more similar outcomes (e.g., earnings or employment status) with respect to people working in different places. This is a well-known issue in labour economics addressed, for example, by Heckman et al. (1997). The authors pointed out that matching methods are far more effective in recovering the parameter of interest when the comparison and treatment group both are drawn from the same local labour markets since both the levels and dynamics of earnings and employment are affected by the conditions of the local labour market in which persons are located. This is confirmed also by Friedlander and Robins (1995) and Bloom et al. (2002) which compare, in the U.S. context, findings for a number of experimental comparison groups with those for non-experimental control groups obtained both from the same and from different states with respect to the program samples. The resulting non-experimental estimates were usually quite different from the experimental estimates derived from the same data when out-of-state comparison groups are used. On the contrary, when comparison samples are drawn in the same state as the program sample the average discrepancy between experimental and non-experimental estimates was smaller. These studies illustrate the risks involved in comparing the behaviour of individuals residing in different geographic areas and can be generalised to all situations where context *matters*, that is where people residing in different areas are subject to different environments, and these are likely to affect the selection and outcomes under study. For example, Aassve and Arpino (2007) analysed the effect of childbearing events on economic wellbeing in rural Vietnam using data from the Vietnamese Living Standard Measurement Survey (VLSMS). The contextual dimension, represented by the community characteristics, plays an important role in this application. The authors can benefit from an important series of information at the community level which is available for the rural sample of the VLSMS (concerning for example health facilities, educational indexes, the presence of roads and other infrastructures). But this is not always the case. For example, community information is unavailable for the urban sample of the VLSMS, likewise for some others surveys of this kind (e.g., the LSMS for Armenia, Bosnia and Herzegovina, Romania, Serbia). Estimates of the fertility effect on wellbeing on this kind of data would be affected by the problem of omitted cluster level variables. In this paper we address the issue of the specification of the propensity score model when an unmeasured context problem is at hand. In the next section we characterise the different situations we can encounter when evaluating treatment effects in multilevel studies, in terms of the assumptions we can make on the assignment mechanism, and we discuss alternative specifications of the propensity score which aim to face a potential unmeasured context problem. We consider typical data structures in large scale surveys, characterised by a relative high number of small clusters as in Aassve and Arpino (2007), where the sample size consists of 2023 households clustered in 120 communities.

3. The specification of the propensity score model for multilevel data

In this section we adapt the general framework outlined in the previous section to a multilevel setting. Let suppose to have a two-level data structure in which N micro units at the first level, indexed by i ($i = 1, 2, \dots, n_j$), are nested in J macro units at the second level, indexed by j ($j = 1, 2, \dots, J$). We can have variables measured both at the first (X) and at the second level (C). With respect to the level of assignment of treatment(s) we can have the following situations:

a) one (or more) treatment(s) assigned at the individual level
b) one (or more) treatment(s) assigned at the second level
c) one (or more) treatment(s) assigned at each level

In this paper we focus on situations of type *a*, where the treatment is assigned at the individual level only, and we distinguish between the following two situations:

1. the treatment is assigned at the individual level but one or more cluster-level characteristics affect both the selection into treatment status and on the potential outcomes. The equivalent randomised experiment is the so-called multi-site experiment, where units are randomly assigned to treatment or control within cells jointly defined by unit and cluster-level characteristics.

2. the treatment is assigned at the individual level but belonging to a cluster instead that to another has an affect the selection (and the potential outcomes). This corresponds to the so-called cluster randomised experiment, where first the level of treatment is randomly assigned to clusters, then within clusters units are randomly assigned to treatment or control on the basis of blocks defined by individual characteristics.

In both situations the cluster effects can operate in two ways. It could be restricted to the fact that the probability to get the treatment changes by clusters (according to clusters characteristics, in the first case, or cluster-belonging, in the second) but, for units with the same individual characteristics the relative percentage of treated units is fixed among clusters. In other words, the effect of individual characteristics on the probability to be selected into treatment is fixed. Alternatively, both the overall probability and the relative risk to get the treatment, depending on the individual characteristics, could change by cluster. In this case, there is interaction between the cluster and the individual effects.

The previous classification is important both conceptually and for the statistical implications. In the first situation what is relevant for the analysis is the knowledge of the clusters characteristics and not cluster-belonging *per se*. Said in other words, we aim to compare treated and controls with both first and second level similar characteristics and we do not need to force that matched units belong to the same cluster. On the contrary, in the second situation, we would compare treated and control units belonging to the same cluster and with similar individual characteristics.

More formally, the previous distinction can be done with respect to the way cluster effects enter in generating the treatment status and potential outcomes and consequently with respect to the unconfoundedness assumption needed to identify causal effects. It turns out natural to specify the unconfoundedness assumption for the two situations, respectively, as follows:

Unconfoundedness assumption – case 1 (A.3): $Y_1, Y_0 \perp D | X, C$;

Unconfoundedness assumption – case 2 (A.4): $Y_1, Y_0 \perp D | X, C_1, C_2, \ldots, C_J$.

Two simple data generation mechanisms that conform to these assumptions are the following:

Data generation model – 1

(Treatment)

$T_{ij}* = X_{ij}\beta + C_j \alpha + \varepsilon_{ij}$

$T_{ij} = \mathrm{I}\,(T_{ij}* > 0)$

(Potential Outcomes)

$Y_{1ij} = X_{ij}\,\delta_1 + C_j\,\theta_1 + \eta_{1ij}$

$Y_{0ij} = X_{ij}\,\delta_0 + C_j\,\theta_0 + \eta_{0ij}$

Data generation model – 2

(Treatment)

$T_{ij}^{*} = X_{ij}\beta + C_{1j}\,\alpha_1 + C_{2j}\,\alpha_2 + \ldots + C_{Jj}\,\alpha_J + \varepsilon_{ij}$

$T_{ij} = \mathrm{I}\,(T_{ij}^{*} > 0)$

(Potential Outcomes)

$Y_{1ij} = X_{ij}\,\delta_1 + C_{1j}\,\theta_{11} + C_{2j}\,\theta_{21} + \ldots + C_{Jj}\,\theta_{J1} + \eta_{1ij}$

$Y_{0ij} = X_{ij}\,\delta_0 + C_{1j}\,\theta_{10} + C_{2j}\,\theta_{20} + \ldots + C_{Jj}\,\theta_{J0} + \eta_{0ij}$

where X and C in both models represent, respectively, a set of first and second level characteristics. In the first model ε is a random error uncorrelated with η_1 and η_0, while η_1 and η_0 are allowed to be correlated. In the second model, C_1, C_2 ,…, C_J are dummies for the J clusters (obviously the overall intercepts are omitted in this case). In both model, we could allow for interactions among cluster and individual effects.

It is important to note that it does not make sense to order assumptions (A.3) and (A.4) with respect to their weakness. They are simply different and can be both more plausible than the other in a given application. However, there are practical implications for the implementation of the propensity score methods. Depending on the assumption we make we would use different propensity score matching strategies, with respect both to the specification of the propensity score and the matching algorithm. Our paper focuses on case (1) since it is the most common in observational studies. In the following we consider that an unmeasured context problem is at hand or, said, in other words that assumption (A.3) does not hold due to the fact that one or more relevant cluster level covariate is unobserved. For simplicity we consider a single cluster level covariate which is unobserved to the researcher. Our aim is to compare different specifications for the propensity score in such a situation. Our interest lies in the ability of the propensity score model to take into account the cluster effects and thus reduce the biasing effect on the ATT due to omission of a relevant macro variable.

Several strategies for PSM implementation can be used in a situation like case 1. Among them we consider the following:

Strategy 1

Propensity score specification (simple single level):

$\pi_i = F(X_i\lambda)$

Propensity score estimates:

$\hat{\pi}_i = F(X_i\hat{\lambda})$

Matching: within clusters

Strategy 2

Propensity score specification (cluster-specific single level):
$\pi_{ij} = F(X_{ij}\lambda_j)$
Propensity score estimates:
$\hat{\pi}_{ij} = F(X_{ij}\hat{\lambda}_j)$
Matching: within clusters

Strategy 3

Propensity score specification (multilevel random intercept):
$\pi_{ij} = F(X_{ij}\lambda + u_j)$
Propensity score estimates (empirical bayes probabilities):
$\hat{\pi}_{ij} = \int F(X_{ij}\hat{\lambda} + \hat{u}_j) \times Posterior\ (u_j \mid y_{1j},..., y_{n_j j})du_j$
Matching: not forced to be within clusters

Strategy 4

Propensity score specification (2-stage):
1st stage: $\pi^1_{ij} = F(X_{ij}\lambda + u_j)$
2nd stage: $\pi^2_{ij} = F(X_{ij}\lambda + \hat{u}_j\omega)$
Propensity score estimates:
$\hat{\pi}_{ij} = F(X_{ij}\hat{\lambda} + \hat{u}_j\hat{\omega})$
Matching: not forced to be within clusters

Strategy 5

Propensity score specification ("dummies model"):
$\pi_{ij} = F(X_{ij}\lambda + C_{1j}\ \gamma_1 + C_{2j}\ \gamma_2 + ... + C_{Jj}\ \gamma_J)$
Propensity score estimates:
$\hat{\pi}_{ij} = F(X_{ij}\hat{\lambda} + C_{1j}\ \hat{\gamma}_1 + C_{2j}\ \hat{\gamma}_2 + ... + C_{Jj}\ \hat{\gamma}_J)$
Matching: not forced to be within clusters

Since our focus is on large scale observational studies (e.g. national surveys) where the typical data structure, characterised by a relatively large number of small clusters (few observations per cluster) makes the implementation of the matching algorithm within clusters, which would solve the omitted variable problem, difficult we consider only strategies 3 to 5.

4. The simulation procedure and results

In this section we introduce the general setup and the results of the first set of simulations we currently have completed.

The setup of the Monte Carlo simulation in this paper builds on the setup used by Zhao (2005). As in Zhao's paper also in our work the focus is on the bias (and the mean squared error) of the ATT estimators. However, while Zhao assesses the robustness of the estimated treatment effect to misspecifications (concerning the error term specification and the included covariates) of the propensity score, we compare alternative specifications of the propensity score in a multilevel setting with unobserved cluster level covariates.
We generate two-level balanced data structures, where the overall sample size, *N*, is determined as the product of the number of clusters, *nc*, and the fixed cluster size, *cs*.

As we said in the previous section, our paper focuses on situations where the data conforms to an unconfoundedness assumption like (A.3). In particular, imposing the condition that both selection into treatment and potential outcomes depend on three first level covariates X_1, X_2 and X_3 and one cluster level covariate, *C*, the unconfoundedness assumption under which our simulation study is carried on is:

Unconfoundedness assumption – special case of (A.3): $Y_1, Y_0 \perp T \mid X_1, X_2, X_3, C.$

In particular and similar to Zhao, we use the following data generation mechanism:

(Treatment)

$$T_{ij}^* = \beta_0 + X_{1ij}\,\beta_1 + X_{2ij}\,\beta_2 + X_{3ij}\,\beta_3 + \alpha\, C_j + \varepsilon_{ij}$$
$$T_{ij} = \mathrm{I}\,(T_{ij}^* > 0)$$

(Potential Outcomes)

$$Y_{1ij} = \delta_{10} + \delta_{11} X_{1ij} + \delta_{12} X_{2ij} + \delta_{13} X_{3ij} + \theta_1 C_j + \eta_{1ij}$$
$$Y_{0ij} = \delta_{00} + \delta_{01} X_{1ij} + \delta_{02} X_{2ij} + \delta_{03} X_{3ij} + \theta_0\ C_j + \eta_{0ij}$$

Both X_1 and the error terms are generated as standard normal variables, while X_2 is generated from a chi-square distribution and X_3 is a mixture of two normal distributions. We allow for correlation between the error terms η_{1ij} and η_{0ij}, but impose the error terms in the outcome equations to be uncorrelated with the error term in the selection equation. This amounts to imposing that the unconfoundedness assumption as specified above is respected. In the simulation procedure we fix parameters β, δ and θ, but allow α to vary. As α increases, that is the cluster effect becomes stronger, the biasing effect of omitting *C* turns out to be more important and we expect that using specifications that accounts for the omitted variable becomes more relevant.

We compare strategies 3, 4 and 5 as outlined in the previous section. As a reference we also estimate two single-level propensity score models, one including and the other excluding the variable *C*. Finally, we estimate a single level propensity score with the inclusion of the cluster means of X_1 as a substitute for *C*. In all cases, the employed matching method is nearest neighbour with replacement combined with a caliper of 0.01.

Apart from the different values employed for the parameter α (0.1; 0.3; 0.5), the several setups considered in the simulations differ in terms of how *C* is generated and the data structure (number of clusters and cluster size). The cluster variable *C* is generated both to be uncorrelated and correlated with the first level covariates. We generated the *C* variable in four ways: a) uncorrelated with *X* and generated by a standard normal distribution; b) uncorrelated with *X* and generated by a chi-square distribution with one degree of freedom; c) uncorrelated with *X* and generated by a mixture of two normal variables; d) correlated with X_1: $C = a + b\overline{X}_1 + error$. When *C* is not normally distributed or is correlated with a first level variable the multilevel models used in strategies 3 and 4 are misspecified. We are interested in the effect of these misspecifications on the estimated ATT. Moreover, when the unobserved cluster variable *C* is correlated with X_1 we could expect that including in the propensity score the cluster means of this variable, $\overline{X}_1$, together with the first level covariates helps in balancing *C*. This idea relates to the so-called second level endogeneity problem in the multilevel modelling literature. This problem arises when in a multilevel regression model we omit a cluster level covariate which is correlated with a first level one. As a consequence, the error term at the cluster level will be correlated with the first level covariate, leading to inconsistent parameter estimates. The problem is circumvented by including the cluster mean of the endogenous first level variable (see e.g., Snijders and Bosker, 1999).

As far as the data structure is concerned, when we allow the number of clusters to vary we fix the cluster size and vice versa. Both increasing the number of clusters (holding the clusters size constant) and increasing the cluster size (holding the number of clusters fix) result in an increase of the sample size and, as a consequence, each estimator should perform better. However, we are interested in comparing the relative performances of the several strategies. In the first set of simulations, whose results are shown in this paper, we fix the cluster size to 20 units and consider four values for the number of clusters (25, 50, 100, 200). For the second set of simulations we plan to develop, we will fix the number of clusters to 50 and consider four values for the clusters size (10, 20, 40, 80). This design also allows comparison of three situations with the same sample size but different data structure (*nc*=50 and *cs*=10 versus *nc*=25 and *cs*=20; *nc*=100 and *cs*=20 versus *nc*=50 and *cs*=40; *nc*=200 and *cs*=20 versus *nc*=50 and *cs*=80).

Tables 1 and 2 summarise the results of the first set of simulations, where the number of clusters changes but the cluster size is hold constant to 20 units. In the tables we report only the results for the two extreme data structures (*nc*=25 and *nc*=200) and for α = 0.5 (highest effect of the cluster-level confounder in the data generating model for the true propensity score). The results for the other cases are qualitatively the same as those reported in the tables. In the simulations, we used a PSM with replacement and caliper = 0.01. In table 1, we generated the cluster-level confounder as a normal variable uncorrelated with the *X*, while in Table 2 we generated this variable to be normal but correlated with X_1. The results we obtained when *C* was generated to have a chi-square or a bi-modal distribution are not shown here. These results, which are consistent with those shown in the paper, are available from the authors upon request.

The following models are compared in the tables:

M1 = single level logit including X_1, X_2, X_3, *C* as covariates
M2 = single level logit including X_1, X_2, X_3 as covariates
M3 = single level logit including X_1, X_2, X_3, $\overline{X}_1$ as covariates
M4 = two-level logit (strategy 3) including X_1, X_2, X_3 as covariates
M5 = two-stage procedure (strategy 4) including X_1, X_2, X_3, $\overline{X}_1$ as covariates
M6 = single level logit including X_1, X_2, X_3, C_1, C_2,…, C_J as covariates

From tables 1 and 2 we can see that the two-stage procedure (M5) and the dummy model (M6) show acceptable bias (measured as absolute standardise bias, ASB[1].) and error (MSE) when compared to the benchmark model, M1. (M1 uses *C* as a covariate in the estimation of the propensity score and corresponds to the case where we have no omitted variables.)

In most cases, the magnitudes of both bias and MSE for models M5 and M6 are comparable to those of M1. These methods perform much better than the single level model which does not take into account the omitted cluster level variable at all (M2). This is because both the two-stage and the dummy model achieve a reasonably good balancing of the omitted variable, *C*, between the treated and control groups, as attested by the ASB calculated after the matching. The inclusion of the cluster mean of the first level variable X_1 (model M3) does not significantly improve the performance of model M2. Finally, a standard two-level logistic regression (M4) shows better bias and MSE with respect to a single level one (M2), but in most cases its performance is significantly worse than the two-stage and the dummy procedures.

The previous results are not affected substantially by the way *C* is generated, by its effect on the selection into treatment (α) or by the number of clusters in the generated data (*nc*).

Between the two-stage and the dummy procedure there is no a clear "winner". In several cases the dummy procedure shows lower bias and MSE; in other cases the bias is lower but the MSE is higher. In most cases, however, there are no huge discrepancies between the two methods.

5. Concluding remarks

In this paper we address the problem related to the bias in the average treatment effect estimated with a propensity score matching procedure in the presence of unobserved higher level covariates. This problem arises in multilevel structured data where the contextual heterogeneity is not fully captured by the observed variables in the data set.

We clarify the assumptions needed to identify causal effects in different multilevel settings. Our focus is on situations where we do not have a different treatment in different clusters, but cluster-level characteristics affect both the probability to take the treatment and the potential outcomes and can be considered as confounding variables, like the individual-level ones. In these cases within-cluster matching is not needed and propensity score specifications that take into account the unobserved cluster-level heterogeneity can be used. Using Monte Carlo simulations we compare the performance (bias and MSE) of multilevel and fixed-effect

[1] The ASB, suggested by Rosenbaum and Rubin (1985), is defined as the absolute difference of sample means in the treated and matched control subsamples as a percentage of the square root of the average of sample variances in both groups. In formula, the ASB is given by:

$$ASB = \left| 100 \frac{\left(\bar{X}_T - \bar{X}_C\right)}{\sqrt{0.5\left(s_T^2 - s_C^2\right)}} \right|$$

where for each covariate $\bar{X}_T$ and $\bar{X}_C$ are the sample means, respectively, in the treated and control group and s_T^2 and s_C^2 are the corresponding sample variances. One possible problem with the standardised bias approach is that one does not have a clear indication for the success of the matching procedure. Again, we compare the ASB for models M2-M6 with the ASB of our benchmark model, M1.

models (single level model with dummies for clusters). Among the multilevel specifications we propose a two-stage procedure that first estimates the contextual effects, as captured by the empirical bayes predictions of the random effects, and then, at the second stage, uses these predictions as a covariate to be used along with the observed potential confounds in the propensity score estimation. We find that the two-stage procedure (strategy M5) and the fixed-effect model (strategy M6) serve quite well the scope of capturing the unobserved heterogeneity. In fact, the bias and MSE for these strategies are comparable with our benchmark, which is represented by the fully specified propensity score procedure (that is, the one that assumes the cluster-level variables are observed). As a confirmation, these strategies achieve a balancing in the unobserved cluster-level variable which is comparable to that obtained with the benchmark model. This result is important for those situations, quite common in large-scale surveys where data are collected only at the lowest levels (for example, individuals and households), while no information is given at the higher levels (for example, communities or provinces).

However, further analyses are needed. We plan to develop simulations with smaller cluster sizes and with unbalanced data structures. We expect that in these cases the dummy model will lose efficiency with respect to the multilevel specifications.

References

Aassve, A. and Arpino B. (2007) Estimation of causal effects of fertility on economic wellbeing: Evidence from rural Vietnam, ISER Working Paper 2007-24. Colchester: University of Essex.

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.

Agresti, A. (2002) *Categorical Data Analysis*, 2nd edition. New Jersey: Wiley.

Bloom, H. S., Michalopoulos, C., Hill, C. J. and Lei, J. (2002) Can non-experimental comparison group methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? MDRC Working Paper on Research Methodology, available at http://www.mdrc.org/ResearchMethodologyPprs.htm.

Blundell, R., Dearden, L. and Sianesi B. (2005) Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS. *Journal of the Royal Statistical Society, Series A*, 168(3), 473-512.

Brand, J. E. and Halaby, C. N. (2006) Regression and Matching Estimates of the Effects of Elite College Attendance on Education and Career Achievement. *Social Science Research*, 35, 749-770.

Bryson, A., Dorsett, R. and Purdon S. (2002) The use of propensity score matching in the evaluation of labour market policies. Working Paper No. 4, Department for Work and Pensions.

Bryson, A. (2002) The Union Membership Wage Premium: An Analysis Using Propensity Score Matching, Discussion Paper No. 530, Centre for Economic Performance, London School of Economics.

Caliendo, M. and Kopeining, S. (2008) Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1), 31-72.

Cox, D. R. (1958) *Planning of Experiments*. New York, Wiley.

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

Dehejia, R., and Wahba, S. (1999) Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94 (448), 1053–1062.

Eren, O. (2007) Measuring the Union-Nonunion Wage Gap Using Propensity Score Matching. *Industrial Relations,* 46(4), 766-780.

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

Friedlander, D. and Robins, P. K. (1995) Evaluating Program Evaluations: New Evidence on Commonly Used Non-experimental Methods, *The American Economic Review*, 85(4), 923-937.

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

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(4), 605-654.

Heckman, J. J., Ichimura, H. and Todd, P. (1998) Matching as an Econometric Evaluation Estimator. *Review of Economic Studies*, 65(2), 261-294.

Hong, G., and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, 101(475), 901-910.

Hox, J. J. (1995) *Applied Multilevel Analysis.* Amsterdam: TT- Publikaties.

Ichino, A., Mealli, F. and Nannicini, T. (2008) From Temporary Help Jobs to Permanent Employment: What Can We Learn from Matching Estimators and their Sensitivity? *Journal of Applied Econometrics*, 23(3), 305-327.

Imbens, G. W. (2004) Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1), 4-30.

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.

Manski C. F. (1990) Nonparametric Bounds on Treatment Effects, *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. F. and Garfinkel, I. (1992) *Evaluating Welfare and Training Programs*, Cambridge, MA: Harvard University Press.

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).

Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rosenbaum P. and Rubin D. (1983b), Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society, Series B*, 45, 212-218.

Rosenbaum, P. R. and Rubin, D. B. (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39(1), 33-38.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 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, 591-593.

Sianesi (2004) An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s. *The Review of Economics and Statistics*, 86(1), 133-155.

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

Zhao, Z. (2005) Sensitivity of Propensity Score Methods to the Specifications. IZA Discussion Paper No. 1873.

Tables

Table 1. Monte Carlo results for various PSM strategies. Variable *nc*, fixed *cs* = 20, α = 0.5. Unobserved cluster covariate (C): normal and uncorrelated with *X*.

	M1	M2	M3	M4	M5	M6
			nc=25 (n=500)			
True ATT	0.906	0.906	0.906	0.906	0.906	0.906
Estimated ATT	0.877	1.667	1.680	0.781	0.858	0.933
Bias	-0.029	0.762	0.773	-0.131	-0.053	0.027
MSE	0.048	0.625	0.640	0.066	0.064	0.054
ASB Before X1	38.997	38.997	38.997	38.997	38.997	38.997
ASB Before X2	13.668	13.668	13.668	13.668	13.668	13.668
ASB Before X3	53.435	53.435	53.435	53.435	53.435	53.435
ASB Before C	100.680	100.680	100.680	100.680	100.680	100.680
ASB After X1	9.737	6.476	6.162	11.789	10.136	9.736
ASB After X2	10.121	6.705	7.788	10.816	10.822	11.244
ASB After X3	7.887	4.627	4.582	14.700	9.341	9.042
ASB After C	5.369	81.224	82.950	15.261	8.059	6.357
			nc=200 (n=4000)			
True ATT	0.908	0.908	0.908	0.908	0.908	0.908
Estimated ATT	0.912	1.670	1.664	0.736	0.867	0.920
Bias	0.004	0.762	0.757	-0.172	-0.041	0.013
MSE	0.008	0.584	0.577	0.041	0.014	0.014
ASB Before X1	41.054	41.054	41.054	41.054	41.054	41.054
ASB Before X2	20.157	20.157	20.157	20.157	20.157	20.157
ASB Before X3	58.777	58.777	58.777	58.777	58.777	58.777
ASB Before C	99.511	99.511	99.511	99.511	99.511	99.511
ASB After X1	3.324	1.912	1.934	9.655	4.631	4.509
ASB After X2	4.430	2.448	2.494	7.343	5.792	5.674
ASB After X3	3.445	1.364	1.453	14.482	5.066	4.181
ASB After C	2.141	80.427	80.114	16.407	4.716	2.872

Notes: Replications:300. ASB = absolute standardised bias. MSE = mean squared error; nc = number of clusters; cs = cluster size, n = sample size; α = coefficient of C in the data gene rating model.

Table 2. Monte Carlo results for various PSM strategies. Variable nc, fixed cs = 20, α = 0.5. Unobserved cluster covariate correlated with X1 (r = 0.63).

	M1	M2	M3	M4	M5	M6
			nc=25 (n=500)			
True ATT	0.908	0.908	0.908	0.908	0.908	0.908
Estimated ATT	0.901	1.647	1.423	0.761	0.914	0.906
Bias	-0.006	0.743	0.517	-0.145	0.005	0.002
MSE	0.050	0.592	0.308	0.077	0.060	0.066
ASB Before X1	51.845	51.845	51.845	51.845	51.845	51.845
ASB Before X2	21.098	21.098	21.098	21.098	21.098	21.098
ASB Before X3	52.926	52.926	52.926	52.926	52.926	52.926
ASB Before C	106.587	106.587	106.587	106.587	106.587	106.587
ASB After X1	9.676	5.354	7.149	12.308	10.850	10.285
ASB After X2	10.266	7.312	7.904	11.410	10.760	11.967
ASB After X3	8.504	5.573	7.196	14.996	9.411	10.469
ASB After C	5.843	80.280	48.028	16.891	7.755	7.383
			nc=200 (n=4000)			
True ATT	0.908	0.908	0.908	0.908	0.908	0.908
Estimated ATT	0.901	1.647	1.423	0.761	0.914	0.906
Bias	-0.006	0.743	0.517	-0.145	0.005	0.002
MSE	0.050	0.592	0.308	0.077	0.060	0.066
ASB Before X1	51.845	51.845	51.845	51.845	51.845	51.845
ASB Before X2	21.098	21.098	21.098	21.098	21.098	21.098
ASB Before X3	52.926	52.926	52.926	52.926	52.926	52.926
ASB Before C	106.587	106.587	106.587	106.587	106.587	106.587
ASB After X1	9.676	5.354	7.149	12.308	10.850	10.285
ASB After X2	10.266	7.312	7.904	11.410	10.760	11.967
ASB After X3	8.504	5.573	7.196	14.996	9.411	10.469
ASB After C	5.843	80.280	48.028	16.891	7.755	7.383

Notes: Replications:300. ASB = absolute standardised bias. MSE = mean squared error; nc = number of clusters; cs = cluster size, n = sample size; α =coefficient of C in the data gene rating model; r = correlation between X1 and C.