Identifying Treatment Effects and Counterfactual Distributions using Data Combination with Unobserved Heterogeneity

Pablo Lavado and Gonzalo Rivera
Universidad del Pacífico

December, 2015

Abstract

This paper considers identification of treatment effects when the outcome variables and covariates are not observed in the same data sets. Ecological inference models, where aggregate outcome information is combined with individual demographic information, are a common example of these situations. In this context, the counterfactual distributions and the treatment effects are not point identified. However, recent results provide bounds to partially identify causal effects. Unlike previous works, this paper adopts the selection on unobservables assumption, which means that randomization of treatment assignments is not achieved until time fixed unobserved heterogeneity is controlled for. Panel data models linear in the unobserved components are considered to achieve identification. To assess the performance of these bounds, this paper provides a simulation exercise.

1 Introduction

The need of combining information present in different data sets is very common in socioeconomic investigation. One data set usually does not contain all the relevant information or all the variables needed by the investigator for several reasons; being high surveying costs one of the most important ones. Nevertheless, the potential advantages of using information from different data sets are considerably large, especially in the impact evaluation field.

In this paper, we consider how to identify counterfactual distributions and treatment effects when outcome and demographic variables are found in different data sets. We obtain identification through the construction of sharp bounds based on previous results developed by Fan *et al.* (2014a, 2014b), who adopted the selection on observables assumption. We relax this last assumption allowing for the possibility that a randomization of treatment assignments is not achieved until unobserved heterogeneity is properly controlled for.

To deal with these unobserved factors we use panel data. Particularly, we work in the context where unobserved heterogeneity remains fixed across time and affects both the potential outcomes and program participation in a linear manner[1]. This way, it is possible to perform some transformation to the model (for example, first differences) that removes the unobserved heterogeneity and allows us to "return" to the standard unconfoundedness assumption, where the bounds developed by Fan *et al.* (2014a, 2014b) are valid.

In the impact evaluation field, many studies have investigated the identification and inference of treatment effects when the outcome and demographic variables are observed in a single data set under the selection on observables assumption (see, for example, Chernozhukov *et al.* 2013; Hirano *et al.* 2003; y Rothe 2012). In this context, as mentioned by Fan *et al.* (2014a), the marginal and conuterfactual distributions of potential outcomes (and, thereby, the treatment effects) are pointly identified.

However, the unconfoundedness assumption is not always the most appropriate. In several situations this assumption is violated, which entails that treatment variables are endogenous due to unobserved heterogeneity and selection bias. Literature has developed several methods to face this problem, being the most common the instrumental variables (Heckman *et al.* 1997) and fixed effect panel data approaches (Lillard y Willis 1978; Hislop 1999; Kahn 2007).

[1] See Klevmarken (1982) or Angrist y Krueger (1995) for consistent estimators using instrumental variables when the relevant variables are in different data sets.

This paper is based on the results obtained by Fan *et al.* (2014a), who adopting the selection on observables assumption develop explicit representations of the marginal and counterfactual distributions via an inverse propensity-score reweighting of the data; and, jointly with the Cambanis-Simons-Stout inequality (see Cambanis *et al.* 1976), obtain sharp bound for the marginal and counterfactual distributions and for treatment effects (particularly average treatment effect – ATE- and average effects of treatment on the treated – ATT). Recent literature has used the idea of bounding distributions for partial identification in the impact evaluation field (see Frank *et al.* 1987; Fan y Park (2012, 2010, 2009); Heckman *et al.* (1997); Fan y Zhu (2009), who adopt the unconfoundedness assumption; and Jun *et al.* (2014), who used panel data to deal with the selection on unobservables assumption). Nevertheless, these previous works assume that all relevant variables can be found in a single data set, an ideal situation but that may not be entirely realistic. In this new context, the ides of bounding distributions can result very useful for identification.

Literature related to data combination, although small, is showing an important growth in the last years. Ridder and Moffitt (2007) discussed the use of Frèchet-Hoeffding inequalities when combining two data sets with no common individuals. Cross and Manski (1999) developed Sharp bounds on a regression of the outcome variable (Y) on two discrete control variables ($X; Z$) when the conditional distributions $F(Y|X)$ and $F(Z|X)$ can be identified from different data sets. Furthermore, Fan *et al.* (2014a) adopt the standard selection on observables assumption of Rosenbaum and Rubin (1983) to partially identify treatment effects using the Cambanis *et al.* (1976) inequality.

The use of different data sets can be applied to a common problem in impact evaluation known as the ecological inference problem. This is a particular case where the goal is to combine an aggregated outcome data set with an individual demographic data set to make inference at individual level. This issue was first addressed by King (1997) y King *et al.* (2001), who deal with this problem when the objective is to describe, predict and make individual inference.

In this context, Corvalan *et al.* (2015) does not recommend aggregating individual data as the information lost through aggregation precludes identification. Fan *et al.* (2014b), using the results developed by Fan *et al.* (2014a) and adopting the selection on observables assumption, proposed bound estimators of treatment effects[2]. As mentioned by the authors, the problem of identification in the ecological inference context is analogous to the identification problem of mean counter-

[2] Furthermore, they show the estimators to be consistent and asymptotically normal.

factual outcomes in a treatment effect model, where the outcome of interest is observed only at an aggregate level, but the conditional covariates are observed at an individual level. This identification method has been used by Corvalán *et al.* (2015) to analyze the effect of the change from compulsory to voluntary voting on turnout in mayoral elections in Chile.

Literature related to data combination with unobserved heterogeneity is much smaller. Imbens and Newey (2003) used control functions to identify and estimate non separable models under the assumption that the endogenous variable and model perturbations are independent conditionally on the control variable. This paper is intended to contribute presenting identification of treatment effects developing sharp bounds under the selection of unobservables assumption when the outcome and the covariates are not observed in a single data set.

This paper is organized as follows. Section II introduces the modelling framework, as well as some examples that explain the utility of data combination under the selection on unobservables assumption. Section III presents the main identification results for the case of a linear model on the unobserved heterogeneity, where it is possible to perform some transformation to the model to return to the selection on observables assumption, where the bounds developed by Fan *et al.* (2014a) are valid. Section IV applies partial identification results from Section III to the ecological inference context. Section V shows a simulation exercise. Section VI concludes.

2 Modelling Framework

The context where we work follows closely the potential outcomes approach developed by Rubin (1974). Let $D_t \in (0,1)$ denote the binary variable that indicates the two possible treatment states at each period. This way, if we denote T_0 as the treatment period, D_t will take the value of zero for $t < T_0$ and the value of one for the treated from $t \geq T_0$[3]. On the other hand, we denote Y_{dt} as the outcome variable for each one of the possible states $d = 0,1$ in period t. Rubin considers these as potential outcomes and that, in practice, it is only possible to observe one of them. The observed and potential outcomes are related through the following equation: $Y_t = D_t Y_{1t} + (1 - D_t)Y_{0t}$ at each period ($t = 1, \dots, T$; where T is the number of periods in the sample). Finally, we denote X_t to the covariates (usually demographic variables) at time t that can potentially affect both D and $(Y_{1t}; Y_{0t})$.

This framework is frequently used in the impact evaluation field, where the main objective is to compare certain characteristics of the potential outcomes distribution. As mentioned, this objec-

[3] Clearly, $D_t = 0$ for all non treated individual for all t.

tive is usually achieved adopting the selection on observables assumption. However, this paper relaxes this assumption and extends the analysis to a context where a randomized treatment assignment is not achieved until both observed and unobserved heterogeneity is controlled for.

We denote the unobserved heterogeneity as η; which we assume constant across time. Following Jun *et al.* (2014), we can think of η in a vectorial form; in the sense that some elements can be excluded of the equation that determines D or Y_{dt} for $d = 0,1$. Formally, we can represent this approach the following way:

$$Y_{dt} = g_d(X_t; \gamma; u_{dt}) \qquad (1)$$
$$D_t = h(X_t; \delta; v_t) \qquad (2)$$

Where the vector η contains the elements of γ (affects only potential outcomes) and δ (affects only treatment participation). This representation is a non-separable static panel model with specific unobserved heterogeneity for potential outcomes and treatment participation. This paper assume a linear model with common unobserved heterogeneity, that is, $\gamma = \delta = \eta$.

As a departure from existing literature, we assume that the variables $(Y_t; D_t; X_t)$ are not observed in a single data set at time t. Instead, we observe two separate data sets at each period: (i) the one that contains outcome variables $(Y_t; D_t)$ –named *outcome data set*– and (ii) the one that contains the covariates or demographic variables $(X_t; D_t)$ –named *covariate data set.* We work under the assumption that panel data is available for both data sets; so individuals within the outcome and covariate data set will be the same at all time, but they do not have common individuals[4]. Finally, we asume that we have balanced panel data, that is, we have information of all relevant variables of each individual for all the periods in the simple. We present some examples showing the usefulness of this approach below.

Example I*:* Long Term Returns to College Attendance (similar to the example given by Fan *et al.* 2014a). The information problem arises when the outcome variable is a long term indicator. For example, the effect of college attendance on life time earnings. Clearly, there exist unobserved factors that affect both treatment participation – college attendance - and the outcome variable - life time incomes – (i.e genetic factors, cognitive and non-cognitive skills, among others). This

[4] This implies that there cannot exist an individual that is in the outcome data set and the covariate data set simultaneously. Potentially, there is the possibility that an individual is present in both surveys even though it is impossible to identify him. For simplicity, we consider that the probability of this happening is extremely close to zero.

unobserved heterogeneity causes that the treatment effect is biased. In this context, the availability of large panel data is useful to deal with these unobserved factors via a fixed effects estimation. This way, one can use administrative data, which contain information of life time earnings, and combine it with other surveys, tipically household surveys, to obtain unbiased treatment effects.

Example II*:* Change in the income distribution across time (adapted from Fan *et al.* 2014a and DiNardo *et al.* 1996). DiNardo *et al.* (1996) compared income level on two different years, being the treatment variable a binary indicator for each year. In their seminal work, all relevant variables were found in a single data set. Nevertheless, there are certain variables of interest (particularly, supply side wage determinants) that cannot be observed in the same data set as wages. In this situation panel data together with our identification results can be used to control for unobserved heterogeneity and bound the treatment effects.

Example III: Effect of smoking on birth weight (adapted from Jun *et al.* 2014). A mother's choice regarding smoking is, in general, correlated with factors that affect whether she carries a healthy life style, which, in turn, impacts directly to her child's birth weight. For this reason, a randomization of treatment assignments is not achieved unless we control for these factors that, in many cases, are practically impossible to measure or observe.
Even though the authors used a single data set, it is plausible to consider that there are some variables of interest in other data sets. For example, some measurements of the mother's habits during pregnancy available at health surveys. Potentially, the outcome variable –birth weight– and covariates could be observed in different data sets if we consider a survey that only gathers information related to the child and another that gathers information related to the mother. Again, the availability of panel data would be extremely useful to deal with the unobserved heterogeneity in this context.

The use of different data sets could also be applied to the ecological inference problem; that is, the case where we combine an aggregate outcome data set with an individual covariate data set to make inference at an individual level. In this context, we can understand treatment as an aggregate event so D_t denotes, for example, two geographical areas at time t. We illustrate this context with an example related to the electoral sphere.

Example IV: The Effect of the change from compulsory to voluntary voting on turnout in Chilean mayoral elections. (see Corvalán *et al.* 2015). In Chile, until 2012, the electoral participation

required people to register in electoral lists, so even though the register was voluntary, once registered, voting was compulsory. Chilean government issued Electoral Law 20568 in January 2012, which established that, from that moment, registration was automatic and voting voluntary. Corvalan *et al.* (2015) used this change of electoral regime to compare voters turnout in mayoral elections in 2012 against 2004; adopting the methodology developed by Fan *et al.* (2014b) under the selection on observables assumption. Their treatment variable was a binary indicator for each year. The authors used as outcome data set the aggregate voting information in each year and, as covariate data set, a socioeconomic survey at individual level[5]. In this context, the availability of panel data for both data sets will be useful to control for potential unobserved heterogeneity such as quality of candidates, which may affect the voting decision.

An important aspect to consider when we deal with panel data is the timing of the surveys. Depending on the assumptions adopted and the type of model we are working with, information requirements are going to change. For example, in some contexts it may not be necessary to have panel data for the covariate data set since identification can be reached with a baseline, that is, with information prior to the treatment. Nevertheless, for more complex models, more information or stronger assumptions will be needed to achieve identification. In this paper, we work with model where it is possible to remove the unobserved heterogeneity performing a transformation to the variables (which is possible in, for example, linear models). For simplicity, we assume the availability of only one survey prior to the treatment, which will be denoted as $t = 0 < T_0$[6].

In the rest of the document, we will adopt the notation used by Fan *et al.* (2014a), who denote $F_{A|B}(.|b)$ as the cumulative distribution function of random variable A conditional to $B = b$. Likewise, we denote $F^{-1}(.)$ as the quantile or inverse function of the distribution function $F(.)$.

3 Identifying Treatment Effects

In this section we present the identification of counterfactual distributions and treatment effects (ATE and ATT) in models where it is possible to remove the unobserved heterogeneity through the transformation of the potential outcomes.

[5] Their information sources were the National Statistic Institution for voting information and the National Household Survey (Casen) for socioeconomic information at individual level.

[6] The extension to cases where there exist more tan one period prior to the treatment is straightforward.

3.1 Assumptions

First, we present the selection on unobservables assumption. It establishes that, at each period, there is not a randomized treatment assignment until both observed and unobserved heterogeneity is properly controlled for.

Assumption III.1 (A0): Let us consider that $(Y_{1t}; Y_{0t}; D, X_t; \eta)$ have a joint distribution for all the periods in the sample. It follows that, for all $x_t \in \mathcal{X}_t$ and $n \in \mathcal{N}$, $(Y_{1t}; Y_{0t})$ is jointly independent from D given $X_t = x_t$ and $\eta = n$.

An example of this type of models where the unobserved heterogeneity can be removed are the linear models. In this section, we adopt a model where the unobserved factors is included linearly in the potential outcomes an treatment participation's equations:

$$Y_{dt} = g_d(X_t) + \gamma + u_{dt} \qquad (3)$$

A common practice to control for the unobserved heterogeneity in linear models is to use the first difference or the within estimator. Clearly, performing any of these transformations we can remove the unobserved factors in equation (3).

For a more general approach, denote $Y_d = (Y_{d0}; Y_{d1}; \dots; Y_{dT})'$ for $d = 0,1$ as the two $Tx1$ vectors that contain the potential outcomes at each period. We consider the transformation proposed by Arellano (2003) of the form KY_d; where K is a $(T-1)xT$ matrix with rank $(T-1)$ such that $K\iota = 0$ and ι is a $Tx1$ vector full of ones. The orthogonality of the matrix K and the vector ι guarantees the removal of the time fixed unobserved heterogeneity, so the transformed variables KY_d no longer depend on η. It is important that any transformation must include the pre-treatment period ($t = 0$) to identify treatment effects. Both the first difference and the within matrix satisfy this requirements. This way, the transformed potential outcomes does not depend on the unobserved heterogeneity.

Regarding treatment participation, let us recall that we are working with only one pre-treatment period. We assume a linear relation between the unobserved heterogeneity and program participation. So, we work with the first difference of D_t to remove these unobserved factors. Henceforth, we can define $D = \Delta D_t$ as the *treatment condition*; as it takes the value of one for the treatment group and zero for the control group. We model this treatment condition as follows:

$$D = h(X_0) + v \qquad (4)$$

Where X_0 represents the pre-treatment observable characteristics. That is, we model the treatment condition only on initial characteristics[7]. The main reason for this modeling framework is to avoid that the covariates could be affected by the treatment. This type of modelling is common in the impact evaluation literature, particularly when working with matching and difference in difference approach (seer Abadie 2005, Lee 2005).

The advantage of working with these transformations is that the new model fulfills the standard unconfoundedness assumption, based on two conditions. The first one refers to the conditional Independence assumption, while the second condition is related to the support of the propensity score. These assumptions are shown below (adapted from Rosenbaum y Rubin 1983; Firpo 2007 and Fan *et al.* 2014a):

Assumption III.I (A1): Consider that $(KY_1; KY_0; D, X)$ have a joint distribution. It follows that, for all $x \in \mathcal{X}$, $(KY_1; KY_0)$ is jointlu independent from D given $X = x$.

Assumption III.2 (A2): For all $x_0 \in \mathcal{X}_0$ it follows that $0 < p(x_0) < 1$, where $p(x_0)$ is the *propensity score* and it is defined as $p(x_0) = \Pr(D = 1|X_0 = x_0)$.

Where $X = (X_1', \dots, X_T')'$ is the matrix that contains all the covariates for all the periods of the sample. The assumption (A1) indicates that, once the potential outcomes are transformed so that they do not depend on the unobserved heterogeneity, these transformations are independent of the treatment participation conditional on the covariates. On the other hand, assumption (A2) implies that the propensity score based on initial characteristics must be different from zero and one for both the treated and non-treated individuals. This means that treatment participation cannot be predicted in a deterministic manner[8]. This way, under (A1) and (A2), the transformed model satisfies the standard selection on observables assumption[9].

To illustrate how these assumptions work, let us consider the simple but very common case where there are only two periods available (before and after treatment) for both the outcome and the

[7] We could model treatment participation as a function of the "t"-th difference of observed characteristics if we assume that these have not been affected by the treatment. The posterior results will not be modified; the only change is the way we estimate the propensity score.

[8] The case where there is more than one pre-treatment period is straightforward. For example, we could obtain the propensity score from a pool regression incorporating all the characteristics prior to the treatment. Alternatively, we could work with a (weighted) average of the propensity score of each pre-treatment period.

[9] A similar idea was developed by Lee (2005: chapter 4.5).

covariate data set; and that the matrix K represents the first difference operator. In this context, the equations (3) y (4) describes the classical difference in difference model. It is well known that, to achieve identification in this model, we must adopt the common trends assumption. This implies that, in the absence of treatment, the outcome variables of both treated and non-treated individual would have followed a common (parallel) trend conditional to the covariates (see Abadie 2005; Lechner 2013)[10]. It is easy to notice that (A1) is a stronger from of this assumption as it implies that, conditional on the covariates, the first difference of the model is independent of treatment participation; which includes mean independence required by the common trend assumption. Therefore, the differentiated model, we reach a randomized treatment assignment by conditioning only on observable characteristics.

3.2 Marginal and Counterfactual Distributions and Treatment Effects

In the context described above, we can apply the results developed by Fan *et al.* (2014a) to identify marginal and counterfactual distributions as well as treatment effects. As mentioned by Fan *et al.*, when all the variables are not available in a single data set, the distribution $F_{Y_{dt}|X_t,D}(y|x,D)$ is no longer identified. This precludes identification of the counterfactual distributions $F_{Y_{1t}}(y)$, $F_{Y_{0t}}(y)$ y $F_{Y_{0t}|D}(y|1)$, as well as all the parameters that are function of these distributions, like treatment effects[11].

To deal with this problem, Fan *et al.* (2014a) used the Cambanis *et al.* (1976) inequality to obtain sharp bounds of the counterfactual distribution and treatment effects (ATE and ATT). They adopt the selection on observables assumption, whereas this paper allows the treatment condition and the potential outcomes to depend on time fixed unobserved heterogeneity.

However, if we perform the orthogonal transformation described in the previous section, the new potential outcomes, KY_d, no longer depend on the unobserved heterogeneity. Thus, we can adopt a procedure similar to the one developed by Fan *et al.* (2014a) and obtain sharp bounds using the transformed model which satisfies the standard unconfoundedness assumption. This way, the distributions of interest are no longer the ones related to the potential outcomes, but to their transformations: $F_{KY_1}(Ky)$, $F_{KY_0}(Ky)$ y $F_{KY_0|D}(Ky|1)$.

[10] Formally, the common trend assumption can be written as $E[\Delta Y_d|X=x,D=1]=E[\Delta Y_d|X=x,D=0]$.

[11] Note that the potential outcome distribution in presence of the treatment conditional to be treated, $F_{Y_{1t}|D}(y|1)$, is identified from the sample.

Following Fan *et al.* (2014a) and Firpo (2007) it is possible to write the distributions of the transformed potential outcomes as functions of the data, particularly, as a function of the inverse of the propensity score:[12]

$$\begin{aligned} F_{KY_1}(Ky) &= E\left[\frac{D}{p(X_0)} I\{KY \leq Ky\}\right] \\ F_{KY_0}(Ky) &= E\left[\frac{1-D}{1-p(X_0)} I\{KY \leq Ky\}\right] \\ F_{KY_0|D}(Ky|1) &= \frac{1}{p_1} E\left[\frac{(1-D)p(X_0)}{1-p(X_0)} I\{KY \leq Ky\}\right] \end{aligned} \tag{5}$$

Where $p_1 = \Pr[D = 1]$. To have a better understanding of these distribution functions, let us consider the case where only two moments are available (before and after the treatment[13]) and, once again, matrix K is the first difference operator. In this situation, the functions $F_{\Delta Y_1}(\Delta y)$, $F_{\Delta Y_0}(\Delta y)$ represents the probability that the potential outcome growth of a treated and non-treated individual be equal to Δy, respectively.

The expressions described in equation (5) cannot be identified from the available data, so, following Fan *et al.* (2014a), it is possible to establish sharp bounds to partially identify these distributions using the Cambanis *et al.* (1976) inequality[14]. These bounds are presented in Theorem III.1[15]:

THEOREM III.1: For $d = 0,1$, we have that $F^I_{KY_d}(Ky) \leq F_{KY_d}(Ky) \leq F^S_{KY_d}(Ky)$, where:

$$F^I_{KY_1}(Ky) = E\left[D \int_0^{F_{KY|D}(Ky|D)} F^{-1}_{W|D}(u|D)du\right]$$

$$F^S_{KY_1}(Ky) = E\left[D \int_{1-F_{KY|D}(Ky|D)}^{1} F^{-1}_{W|D}(u|D)du\right]$$

$$F^I_{KY_0}(Ky) = E\left[(1-D) \int_0^{F_{KY|D}(Ky|D)} F^{-1}_{V|D}(u|D)du\right]$$

$$F^S_{KY_0}(Ky) = E\left[(1-D) \int_{1-F_{KY|D}(Ky|D)}^{1} F^{-1}_{V|D}(u|D)du\right]$$

[12] The proof is shown in Appendix 1.

[13] Note that the case where we analyze two periods does not limit the analysis to situations where $T = 2$. We could think in a context where there are different horizons depending on whether one is interested in short, median or long term impacts. Alternatively, we could think of a context where we have several outcomes and each one measures a different dimension with different flowering periods.

[14] The Cambanis-Simons-Stout inequality is given in Appendix 2.

[15] Appendix 3 shows a detailed proof.

Furthermore, we have that $F_{KY_1|D}(Ky|1) = E[DI\{KY \leq Ky\}]/p_1$ is already identified, whereas $F_{KY_0|D}(Ky|1)$ is partially identified through: $F^I_{KY_0|D}(Ky|1) \leq F_{KY_0|D}(Ky|1) \leq F^S_{KY_0|D}(Ky|1)$, where:

$$F^I_{KY_0|D}(Ky|1) = \frac{1}{p_1} E\left[(1-D)\int_0^{F_{KY|D}(Ky|D)} F^{-1}_{V/W|D}(u|D)du\right]$$

$$F^S_{KY_0|D}(Ky|1) = \frac{1}{p_1} E\left[(1-D)\int_{1-F_{KY|D}(Ky|D)}^{1} F^{-1}_{V/W|D}(u|D)du\right]$$

Where $W = 1/p(x_0)$ and $V = 1/\big(1-p(x_0)\big)$, and it is assumed that the variances of W, V and V/W are finite. In this context, the bounds are finite and sharp.

Regarding treatment effects, let us denote $\tau = KY_1 - KY_0$ as the individual treatment effect, so we can define the average treatment effect (ATE) and the average treatment effect on the treated (ATT) as $\mu_{ATE} = E[\tau]$ and $\mu_{ATT} = E[\tau|D=1]$, respectively. The sharp bounds of these effects are obtained from Theorem 3.2 of Fan *et al.* (2014a)[16] applied to the transformed model (the case where $g\big(\tilde{Y}_d\big) = \tilde{Y}_d$):

$$\mu_1^L - \mu_0^U \leq \mu_{ATE} \leq \mu_1^U - \mu_0^L$$

$$\frac{E[D(KY)]}{p_1} - \mu^U_{0|1} \leq \mu_{ATT} \leq \frac{E[D(KY)]}{p_1} - \mu^L_{0|1}$$

The advantage of this result is that the bounds can be identified from the available data. Indeed, $F_{KY|D}(Ky|D)$ can be identified from the outcome data set, while the distributions $F_{W|D}(.|D)$, $F_{V|D}(.|D)$ and $F_{V/W|D}(.|D)$ can be identified from the covariate data set. Furthermore, these bounds are considerably narrower than the ones developed by Manski (1990), as shown by Fan *et al.* (2014b).

The identification source of the treatment effects allows us to extract some interesting conclusions regarding the information requirements. Under the selection on time fixed unobservables assumption, we only need two periods from the outcome and covariate data set to estimate sharp bounds for the counterfactual distributions and treatment effects, applying the difference in dif-

[16] The application of Theorem 3.2 of Fan *et al.* (2014a) to the transformed model is described in Appendix 4 (consider the case where $\tilde{Y} = KY$).

ference approach. Moreover, if we model the treatment condition as a function of only pre-treatment characteristics, we would only need panel data for the outcome data set, it suffices with a baseline for the covariate data set. If more periods are available, we could perform more efficient transformations such as the one done by the within group operator.

4 Identifying Treatment Effects with Ecological Inference Data

In this section, we apply the results obtained in section 3 to the case where ecological inference data is available, that is, a context where we seek to combine an aggregate outcome data set with an individual covariate data set to make individual inference. For this purpose, we use simple adaptations of the bounds obtained by Fan *et al.* (2014b) and Corvalan *et al.* (2015) to partially identify the treatment effect. For the ecological inference model, Fan *et al.* (2014b) states that the identification problem in the ecological inference context is analogous to the identification problem of mean counterfactual outcomes in a treatment effect model, where the outcome of interest is observed only at an aggregate level, but the conditional covariates are observed at an individual level.

As Corvalan *et al.* (2015) and Fan *et al.* (2014b), we analyze the case where both the potential and outcomes have a binary behavior. Adopting the selection on observables assumption, the treatment effects can be written as:

$$\mu_{ATE} = E[Y_1 - Y_0] = \Pr[Y = 1] - \Pr[Y = 0]$$
$$\mu_{ATT} = E[Y_1 - Y_0 | D = 1] = \Pr[Y = 1 | D = 1] - \Pr[Y = 0 | D = 1]$$

To estimate these treatment effects, it is only necessary to obtain sharp bounds of these counterfactual means (probabilities). Corvalan *et al.* (2015: Theorem I) and Fan *et al.* (2014b: section 4) obtain this bounds as a particular case of Theorem 3.2 of Fan *et al.* (2014a); as well as plug-in estimators of these bounds.

As in previous section, we could apply the results obtained by Corvalan *et al.* (2015) and Fan *et al.* (2014b) to the transformed model KY_d, which depend only on observed covariates. We consider the particular though very common situation where only two periods are available so the matrix K is the first difference operator[17]. In this context, the treatment effects are

[17] Models where there are more than two periods of time and/or another transformation is performed are straightforward.

$$\mu_{ATE} = E[\Delta Y_1 - \Delta Y_0] = \mathrm{E}[\Delta Y_1] - \mathrm{E}[\Delta Y_0]$$
$$\mu_{ATT} = E[\Delta Y_1 - \Delta Y_0 | D = 1] = \mathrm{E}[\Delta Y_1 | D = 1] - \mathrm{E}[\Delta Y_0 | D = 1]$$

Unlike the case analyzed by Corvalan *et al.* (2015) and Fan *et al.* (2014b), the treatment effects does not reduces to the difference of probabilities as the first difference is no longer a binary variable. Indeed, these transformed variables can take the values of one, minus ones or zero depending on whether the evolution of the potential outcomes is positive, negative or null; respectively.

To solve this problem, we present two alternatives. The first one consists to return to the binary case and only analyze the situation where the potential outcomes has grown (or dropped). This way, we can define an alternative variable that takes the value of one if the growth has been positive (or negative) and zero in any other case. In this context, we estimate the bounds for the counterfactual probabilities as Corvalán *et al.* (2015) and Fan *et al.* (2014b). The second alternative, a little more complex, is to extend the authors results to the case where the outcome is still discrete but can take more than three different values. These two options are analyzed below.

4.1 Dichotomizing potential outcomes growth

Without loss of generality, let us consider the case where the interest is to analyze whether the outcome has positively evolved due to the presence of the treatment. In this context, we can define an auxiliary potential outcome, $\tilde{Y}_d$, as:

$$\tilde{Y}_d = \begin{cases} 1, & \text{if } \Delta Y_d = 1 \\ 0, & \text{if } \Delta Y_d = 0 \; or \; \Delta Y_1 = -1 \end{cases}$$

The observed auxiliary outcome, $\tilde{Y}$, relates the potential outcomes with the following equation: $\tilde{Y} = D\tilde{Y}_1 + (1 - D)\tilde{Y}_0$. The treatment effects are:

$$\mu_{ATE} = E[\tilde{Y}_1 - \tilde{Y}_0] = \Pr[\tilde{Y} = 1] - \Pr[\tilde{Y} = 0]$$
$$\mu_{ATT} = E[\tilde{Y}_1 - \tilde{Y}_0 | D = 1] = \Pr[\tilde{Y} = 1 | D = 1] - \Pr[\tilde{Y} = 0 | D = 1]$$

This way, it is posible to identify treatment effects using sharp bound on the counterfactuals means of the auxiliary variables as Corvalán *et al.* (2015) and Fan *et al.* (2014b). For this purpose, we denote:

$$p_{00} = \Pr[\tilde{Y} = 0|D = 0] = \Pr[\Delta Y_d = 0;\ \Delta Y_1 = -1|D = 0]$$
$$p_{01} = \Pr[\tilde{Y} = 0|D = 1] = \Pr[\Delta Y_d = 0;\ \Delta Y_1 = -1|D = 1]$$

Using these definitions, we can use the bounds developed by Fan *et al.* (2014b) by applying Theorem 3.2 of Fan *et al.* (2014a).

Theorem IV.1: Assuming that $Var(W) < \infty$ and $Var(V) < \infty$, it follows that:

$$L_1^I - L_0^S \leq \mu_{ATE} \leq L_1^S - L_0^I$$
$$\frac{p_{11}}{p_1} - L_{0|1}^S \leq \mu_{ATT} \leq \frac{p_{11}}{p_1} - L_{0|1}^I$$

where:

$$L_1^I = p_1 \int_0^{1-p_{01}} F_{W|D}^{-1}(u|1)du$$
$$L_1^S = p_1 \int_{p_{01}}^1 F_{W|D}^{-1}(u|1)du$$
$$L_0^I = (1-p_1) \int_0^{1-p_{00}} F_{V|D}^{-1}(u|0)du$$
$$L_0^S = (1-p_1) \int_{p_{00}}^1 F_{V|D}^{-1}(u|0)du$$
$$L_{0|1}^I = \frac{(1-p_1)}{p_1} \int_0^{1-p_{00}} F_{V/W|D}^{-1}(u|0)du$$
$$L_{0|1}^S = \frac{(1-p_1)}{p_1} \int_{p_{00}}^1 F_{V/W|D}^{-1}(u|0)du$$

For the estimation of these bounds we can use plug-in estimators analogous to the ones developed by Fan *et al.* (2014b) and Corvalan *et al.* (2015). Let us assume that the outcome data set contains M regions, whereas the covariate data set contains N individuals in both periods. From the outcome data set we can obtain estimators for the sample proportions $\hat{p}_1$, $\hat{p}_{11}$, $\hat{p}_{00}$ and $\hat{p}_{01}$ as proposed by Corvalan *et al.* (2015):

$$\hat{p}_1 = M^{-1} \sum_{i=1}^{M} I\{D_i = 1\}$$
$$\hat{p}_{11} = (M\hat{p}_1)^{-1} \sum_{i=1}^{M} I\{\tilde{Y}_i = 1, D_i = 1\}$$

$$\hat{p}_{00} = (M\hat{p}_1)^{-1} \sum_{i=1}^{M} I\{\tilde{Y}_i = 0, D_i = 0\}$$

$$\hat{p}_{01} = (M\hat{p}_1)^{-1} \sum_{i=1}^{M} I\{\tilde{Y}_i = 0, D_i = 1\}$$

Regarding the quantile functions $F^{-1}_{W|D}(u|0)$, $F^{-1}_{V|D}(u|0)$ y $F^{-1}_{V/W|D}(u|0)$, we can obtain them using the same estimators developed by Fan *et al.* (2014b) and Corvalan *et al.* (2015), with the only difference that our estimators uses the propensity score based on pre-treatment characteristics[18]. Finally, we can obtain estimators of the bounds proposed in Theorem IV.1 by integrating numerically in the established interval.

4.2 Using the potential outcomes growth

The second alternative consists in working directly with the first difference of the potential outcomes. Therefore, it would be very useful to extend the result proposed by Fan *et al.* (2014b) to the case where the outcome variable is still discrete but can take more than two different values.

Consider a general situation, where the potential outcomes, denoted by $\tilde{Y}_d$ can take G distinct values $(\tilde{Y}_1, \dots, \tilde{Y}_G)$; upwardly ordered: $\tilde{Y}_1 < \tilde{Y}_2 < \cdots < \tilde{Y}_G$. To obtain sharp bounds on treatment effects, we can apply once again Theorem 3.2 of Fan *et al.* (2014a) [19] in the particular case where $g(\tilde{Y}_d) = \tilde{Y}_d$. Noting that:

$$F^{-1}_{\tilde{Y}|D}(u|d) = \begin{cases} \tilde{Y}_1 & u \in [0; p_{1d}[\\ \tilde{Y}_2 & u \in [p_{1d}; p_{2d}[\\ \vdots & \vdots \\ \tilde{Y}_G & u \in [p_{G-1d}; 1] \end{cases}$$

$$F^{-1}_{\tilde{Y}|D}(1-u|d) = \begin{cases} \tilde{Y}_1 & u \in [1-p_{1d}; 1[\\ \tilde{Y}_2 & u \in [1-p_{2d}; 1-p_{1d}[\\ \vdots & \vdots \\ \tilde{Y}_G & u \in [0; 1-p_{G-1d}] \end{cases}$$

where, for $d = 0{,}1$ and $j = 1{,}2, \dots, G-1$ we denote $p_{jd} = \Pr[Y = Y_j | D = d]$, sharp bounds takes the form described in Theorem IV.2

[18] Appendix 5 shows in detail the quantile functions estimators proposed by Fan *et al.* (2014b) applied to our context.

[19] Theorem 3.2 of Fan *et al.* (2014a) is shown in Appendix 4.

Theorem IV.2: Assuming that $Var(W) < \infty$ and $Var(V) < \infty$, it follows that:

$$C_1^I - C_0^S \leq \mu_{ATE} \leq C_1^S - C_0^I$$

$$\frac{C_1}{p_1} - C_{0|1}^S \leq \mu_{ATT} \leq \frac{C_1}{p_1} - C_{0|1}^I$$

where:

$$C_1 = E[D\tilde{Y}] = \sum_{i=1}^{G} \tilde{Y}_i p_{i1}^*$$

$$C_1^I = p_1 \left[\int_{1-p_{11}}^{1} \tilde{Y}_1 F_{W|D}^{-1}(u|1)du + \int_{1-p_{21}}^{1-p_{11}} \tilde{Y}_2 F_{W|D}^{-1}(u|1)du + \cdots + \int_{0}^{1-p_{G-11}} \tilde{Y}_G F_{W|D}^{-1}(u|1)du \right]$$

$$C_1^S = p_1 \left[\int_{0}^{p_{11}} \tilde{Y}_1 F_{W|D}^{-1}(u|1)du + \int_{p_{11}}^{p_{21}} \tilde{Y}_2 F_{W|D}^{-1}(u|1)du + \cdots + \int_{p_{G-11}}^{1} \tilde{Y}_G F_{W|D}^{-1}(u|1)du \right]$$

$$C_0^I = (1-p_1) \left[\int_{1-p_{10}}^{1} \tilde{Y}_1 F_{V|D}^{-1}(u|0)du + \int_{1-p_{20}}^{1-p_{10}} \tilde{Y}_2 F_{V|D}^{-1}(u|0)du + \cdots + \int_{0}^{1-p_{G-10}} \tilde{Y}_G F_{V|D}^{-1}(u|0)du \right]$$

$$C_0^S = (1-p_1) \left[\int_{0}^{p_{10}} \tilde{Y}_1 F_{V|D}^{-1}(u|0)du + \int_{p_{10}}^{p_{20}} \tilde{Y}_2 F_{V|D}^{-1}(u|0)du + \cdots + \int_{p_{G-10}}^{1} \tilde{Y}_G F_{V|D}^{-1}(u|0)du \right]$$

$$C_{0|1}^I = \frac{(1-p_1)}{p_1} \left[\int_{1-p_{10}}^{1} \tilde{Y}_1 F_{V/W|D}^{-1}(u|0)du + \cdots + \int_{0}^{1-p_{G-10}} \tilde{Y}_G F_{V/W|D}^{-1}(u|0)du \right]$$

$$C_{0|1}^S = \frac{(1-p_1)}{p_1} \left[\int_{0}^{p_{10}} \tilde{Y}_1 F_{V/W|D}^{-1}(u|0)du + \cdots + \int_{p_{G-10}}^{1} \tilde{Y}_G F_{V/W|D}^{-1}(u|0)du \right]$$

where, for $d = 0,1$ and $j = 1,2,\ldots,G-1$ we denote $p_{j1}^* = \Pr[Y = Y_j; D = 1]$.

It is posible to apply the results described in Theorem IV.2 to the first difference model, where the transformed potential outcome can take up to three different values: one, minus one and zero. If we define $\tilde{Y}_d = \Delta Y_d$ and the quantile function as:

$$F_{\Delta Y|D}^{-1}(u|d) = \begin{cases} -1, & u \in [0; p_{\Delta -1d}[\\ 0, & u \in [p_{\Delta -1d}; p_{\Delta 0d}[\\ 1, & u \in [p_{\Delta 0d}; 1] \end{cases}$$

The bounds for the first difference model are shown in Theorem IV.3.

Theorem IV.3: Assuming that $Var(W) < \infty$ and $Var(V) < \infty$, it follows that:

$$C_1^I - C_0^S \leq \mu_{ATE} \leq C_1^S - C_0^I$$

$$\frac{C_1}{p_1} - C_{0|1}^S \leq \mu_{ATT} \leq \frac{C_1}{p_1} - C_{0|1}^I$$

where:

$$C_1 = E[D\Delta Y] = p_{\Delta 11}^* - p_{\Delta -11}^*$$

$$C_1^I = p_1 \left[\int_0^{1-p_{\Delta 01}} F_{W|D}^{-1}(u|1)du - \int_{1-p_{\Delta -11}}^1 F_{W|D}^{-1}(u|1)du \right]$$

$$C_1^S = p_1 \left[\int_{p_{\Delta 01}}^1 F_{W|D}^{-1}(u|1)du - \int_0^{p_{\Delta -11}} F_{W|D}^{-1}(u|1)du \right]$$

$$C_0^I = (1-p_1) \left[\int_0^{1-p_{\Delta 00}} F_{V|D}^{-1}(u|0)du - \int_{1-p_{\Delta -10}}^1 F_{V|D}^{-1}(u|0)du \right]$$

$$C_0^S = (1-p_1) \left[\int_{p_{\Delta 00}}^1 F_{V|D}^{-1}(u|0)du - \int_0^{p_{\Delta -10}} F_{V|D}^{-1}(u|0)du \right]$$

$$C_{0|1}^I = \frac{(1-p_1)}{p_1} \left[\int_0^{1-p_{\Delta 00}} F_{V/W|D}^{-1}(u|0)du - \int_{1-p_{\Delta -10}}^1 F_{V/W|D}^{-1}(u|0)du \right]$$

$$C_{0|1}^S = \frac{(1-p_1)}{p_1} \left[\int_{p_{\Delta 00}}^1 F_{V/W|D}^{-1}(u|0)du - \int_0^{p_{\Delta -10}} F_{V/W|D}^{-1}(u|0)du \right]$$

where, for $d = 0,1$, we denote $p_{\Delta 11}^* = \Pr[\Delta Y = 1; D = 1]$, $p_{\Delta -11}^* = \Pr[\Delta Y = -1; D = 1]$, $p_{\Delta 0d} = \Pr[\Delta Y = 0 | D = d]$ and $p_{\Delta -1d} = \Pr[\Delta Y = -1 | D = d]$.

To obtain estimators for these bounds, we adopt the ones suggested by Corvalan *et al.* (2015) and Fan *et al.* (2014b), with the only difference that are applied to the model in first differences. Once more, let us assume that the outcome data set contains M regions whereas the covariate data set contains N individuals both before and after treatment. From the outcome data set we can obtain plug-in estimators of the sample proportions $\hat{p}_1$, $\hat{p}_{\Delta 11}^*$, $\hat{p}_{\Delta -11}^*$, $\hat{p}_{\Delta -10}$, $\hat{p}_{\Delta -11}$, $\hat{p}_{\Delta 00}$ and $\hat{p}_{\Delta 01}$:

$$\hat{p}_1 = \frac{1}{M} \sum_{i=1}^M I\{D_i = 1\}$$

$$\hat{p}_{\Delta 11}^* = \frac{1}{M} \sum_{i=1}^M I\{\Delta Y_i = 1, D_i = 1\}$$

$$\hat{p}_{\Delta -11}^* = \frac{1}{M} \sum_{i=1}^M I\{\Delta Y_i = -1, D_i = 1\}$$

$$\hat{p}_{\Delta 01} = \frac{1}{M\hat{p}_1}\sum_{i=1}^{M} I\{\Delta Y_i = 0, D_i = 1\}$$

$$\hat{p}_{\Delta -11} = \frac{1}{M\hat{p}_1}\sum_{i=1}^{M} I\{\Delta Y_i = -1, D_i = 1\}$$

$$\hat{p}_{\Delta 00} = \frac{1}{M(1-\hat{p}_1)}\sum_{i=1}^{M} I\{\Delta Y_i = 0, D_i = 0\}$$

$$\hat{p}_{\Delta -10} = \frac{1}{M(1-\hat{p}_1)}\sum_{i=1}^{M} I\{\Delta Y_i = -1, D_i = 0\}$$

The estimators for the quantile functions $F^{-1}_{W|D}(u|0)$, $F^{-1}_{V|D}(u|0)$ y $F^{-1}_{V/W|D}(u|0)$ can be obtained in the same fashion as in the case where we are dichotomizing the potential outcomes growth, described in Appendix 5. These estimators are similar to the ones used by Fan *et al.* (2014b) and Corvalan *et al.* (2015), but our estimators uses the propensity score based on pre-treatment characteristics. Finally, we can obtain estimators of the bounds proposed in Theorem IV.1 by integrating numerically in the established interval.

V Simulation

To assess the performance of the bounds, we make a simulation exercise. Consider the following two period model $t = 0,1$:

$$Y^*_{1t} = \alpha_1 X_t + \gamma + u_{1t}; \quad Y^*_{0t} = \alpha_0 X_t + \gamma + u_{0t}$$
$$Y_{1t} = I\{\alpha_1 X_t + \gamma + u_1 \geq 0\}; \quad Y_{0t} = I\{\alpha_0 X_t + \gamma + u_0 \geq 0\}$$
$$D = I\{\delta X_0 - \varepsilon \geq 0\}$$

Where $(X_1; X_0; u_{11}; u_{10}; u_{01}; u_{00}; \gamma; \varepsilon) \sim N(0; I_8)$.

This model considers that treatment participation depends only on pre-treatment characteristics. Furthermore, as in the ecological inference context, we will evaluate the case where the potential outcomes are affected linearly by the unobserved heterogeneity.

The main objective is to compare the bounds of the policy parameters of interest, ATE and ATT, proposed in Section IV under the selection on unobservables with the bounds developed by Fan *et*

al. (2014a, 2014b); who adopt the unconfoundedness assumption. Likewise, we compare thess bounds with the difference estimator given by:

$$\frac{\sum_{i=1}^{M} Y_i D_i}{\sum_{i=1}^{M} D_i} - \frac{\sum_{i=1}^{M} Y_i (1 - D_i)}{\sum_{i=1}^{M} (1 - D_i)}$$

Let us recall that this estimator is consistent under the assumption that the simple is completely random; that is, the potential outcomes are independent from the treatment participation. Furthermore, under this assumption, the treatment effects ATE and ATT are the same.

Regarding the simulation, we expect that both the difference estimator and the bounds developed by Fan *et al.* (2014a, 2014b) are biased due to the presence of unobserved heterogeneity. This bias shall not be present in our bounds as they work on the transformed model which has already removed the unobserved heterogeneity. For this exercise, we use as matrix K the first difference estimator, so the bounds used are the ones shown in Theorem IV.3[20].

Table 1 shows the results for the Monte Carlo simulation. The true values of the treatment effects ATE and ATT were computed directly from the simulated data.

[20] To integrate numerically, we used Simpson rule as we were working with discrete functions.

Table 1: Bounds Performance

	True Value	OLS (Y vs D)	Observables (Fan *et al.)*		Unobservables	
			Lower	Upper	Lower	Upper
ATE	-0.018	0.0064	-0.0172	0.0383	-0.0524	0.0487
		(0.0216)	(0.0200)	(0.0195)	(0.0139)	(0.0154)
ATT	0.023	0.0064	-0.0242	0.1311	-0.0834	0.0901
		(0.0216)	(0.0201)	(0.0240)	(0.0315)	(0.0296)

Note: True values were computed directly from the simulated data. The propensity-score was obtained through a probit model. We report the average of 100 repetitions. The sample size was of 2000 observations (1000 per period). Standard errors (across repetitions) are shown in parenthesis. Values: $\delta = 0.5,\ \alpha_1 = 3,\ \alpha_0 = -3$

In the presence of time fixed unobserved heterogeneity, both the difference estimators and the bounds developed by Fan *et al.* (2014a, 2014b) are biased. The former reports a treatment effect extremely close to zero, underestimating the ATE and ATT. On the other hand, the latter seems to overestimate the treatment effects. For the ATT, even though the true value lies within their interval, it is really close to the lower bound; whereas the bounds failed to correctly identify the true value. This problem disappears when we use our bounds develop the Theorem IV.3. Note that, for both the ATE and ATT, the interval midpoint seems to be fairly close to the true value.

VI Conclusions

The need to combine information from different data sets to model causal effects is very common in social sciences. Some potential uses of data combination could be analyzing long term returns of college attendance, modeling electoral behavior (as Corvalan *et al.* 2015), comparing the effects of regional policies, estimating the effects of internal and external wars on individual indicators (such as health, education, etc.), among others. Adopting the selection on observables assumption is possible to partially identify treatment effects using the bounds developed by Fan *et al.* (2014a, 2014b). However, this assumption is not always appropriate. There are many cases where this assumption is violated, which causes the treatment variables are endogenous due to the presence of unobserved heterogeneity. In these cases, an alternative identification source is needed.

In this paper we consider the identification of counterfactual distributions and treatment effects when the outcome variable and covariates are in different data sets under the selection on unobservables assumption. For that purpose, based on the results developed by Fan *et al.* (2014a, 2014b), who obtain sharp bounds to identify counterfactual distributions and treatment effects

(ATE and ATT) under the unconfoundedness assumption. Working with a model based on linear time fixed unobserved heterogeneity, we show that we can apply an analogous procedure to the one proposed by Fan *et al.* (2014a, 2014b) using an orthogonal transformation to the model; so that we can return to the selection on observables assumption. As a particular case, we consider the ecological inference problem, where we combine an aggregated outcome data set with an individual demographic data set to make inference at individual level.

The next step is to search for alternative methods of identification of treatment effects in models based on different forms of unobserved heterogeneity; for example, dynamic models or models that include individual rather than time fixed unobserved heterogeneity. Furthermore, we could consider more flexible specifications allowing for non-linearities in the unobserved heterogeneity. The identification source in this paper is the liner relation between the unobserved variables and the potential outcomes. Nevertheless, this assumption may be too restrictive, so the development of new identification methods robust to different functional forms shall be useful in these contexts.

A possibility is to use alternative bounds available in the literature such as the ones developed by Manski (1990) or the Frèchet-Hoeffding inequalities. Nonetheless, the intervals obtained from these bounds are usually uninformative, so it would be useful to develop alternative forms of identification. This task is very challenging so it is left as future research.

VII References

Abadie, A. (2005). "Semiparametric Difference-in-Difference Estimators". *Review of Economic Studies* 72, 1-19.

Angrist, J. D. and Krueger A. B. (1995). "Split-Sample Instrumental Variable Estimates of the Return to Schooling". *Journal of Business & Economic Statistics*, 13(2), 225-235.

Arellano, M. (2003). *Panel Data Econometrics.* 2da ed. España: Centro de Estudios Monetarios y Financieros (CEMFI).

Cambanis, S.; Simons, G.; and Stout, W. (1976). "Inequalities for Ek(X,Y) When the Marginals Are Fixed". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 36, 285-294. [812, 816].

Chernozhukov, V.; Fernández-Val, I.; and Melly, B. (2013). "Inference on Counterfactual Distributions". *Econometrica*, 81 (6), 2205-2268.

Corvalán, A.; Melo, E.; Sherman, R.; and Shum, M. (2015). *Bounding Causal Effects in Ecological Inference Problems*. Working Paper. California Institute of Technology.

Cross, P. J., and Manski, C. F. (1999). "Regressions, Short and Long". *Econometrica*, 70 (1), 357-368.

DiNardo, J.; Fortin, N.; and Lemieux, T. (1996). "Labor Market Institutions and the Distribution of Wages 1973-1992: A Semiparametric Approach". *Econometrica*, 64, 1001-1044.

Fan, Y.; Sherman, R.; and Shum, M. (2014a). "Identifying Treatment Effects Under Data Combination". *Econometrica*, 82 (2), 811-822.

Fan, Y.; Sherman, R.; and Shum, M. (2014b). *Estimation and Inference in an ecological inference model*. Working Paper. California Institute of Technology.

Fan, Y.; and Park, S. (2012). "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments". *Journal of Econometrics*, 167, 330-344.

Fan, Y.; and Park, S. (2010). "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference". *Econometric Theory*, 26, 931-951.

Fan, Y.; and Park, S. (2009). *Partial Identification of the Distribution of Treatment Effects and Confidence Sets. Advances in Econometrics: Nonparametric Econometric Methods*. Bingley, U.K.: Emerald Group.

Fan, Y.; and Zhu, D. (2009). *Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes.* Working Paper. Department of Economics, Vanderbilt University.

Firpo, S. (2007). "Efficient Semiparametric Estimation of Quantile Treatment Effects". *Econometrica* 75, 259-276.

Frank, M. J.; Nelsen, R. B.; and Schweizer, B. (1987), "Best-Possible Bounds on the Distribution of a Sum–a Problem of Kolmogorov". *Probability Theory and Related Fields* 74, 199-211.

Heckman, J.; Smith, J.; and Clements, N. (1997). "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts". *Review of Economic Studies*, 64, 487–535.

Hirano, K.; Imbens, G. W.; and Ridder, G. (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score". *Econometrica*, 71, 1161-1189.

Hislop, D. R. (1999). "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Woman". *Econometrica*, 67 (6), 1255-1294.

Imbens, G. W.; and Wooldridge, J. (2009). "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature*, 47 (1), 5-86.

Imbens, G.; and Newey, W. (2003). *Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity*. Manuscrito. UC Berkeley: Departamento de Economía.

Jun S. J.; Lee, Y.; and Shin, Y. (2014). *Treatment Effects with Unobserved Heterogeneity: A Set Identificacion Approach*. Pennsilvania: Pennsilvania State University.

Kahn, L. M. (2007). "The Impact of Employment Protection Mandates on Demographic Temporary Employment Patterns: International Microeconomic Evidence". *The Economic Journal*, 117 (521): F333-F56.

King, G.; Rosen, O.; and Tanner, M. (2001). *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Klevmarken, N. A. (1982). *Missing Variables and Two-Stage Least-squares Estimation From More Than One Dataset*. En 1981 Proceedings of the American Statistical Association, Business and Economic Statistics Section, 156–161.

Lechner, M. (2013). *Treatment Effects and Panel Data.* Discussion Paper. Department of Economics, Universität St.Gallen.

Lee, M. J. (2005). *Micro-Econometrics for Policy, Program, and Treatment Effects.* Oxford: Oxford University Press.

Lillard, L. A.; and Willis, R. J. (1978). "Dynamic Aspects of Earning Mobility". *Econometrica*, 46 (5), 985-1012.

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

Ridder, G.; and Moffitt R. (2007). "Econometrics of Data Combination". *The Handbook of Econometrics*, Vol 6B, Chapter 75. Nueva York: North-Holland.

Rosenbaum, P.; and Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*, vol. 70, N° 1, 4155.

Rothe, C. (2012). "Partial Distributional Policy Effects". *Econometrica*, 80, 2269-2301.

Rubin, D. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". *Journal of Educational Psychology*, 66, 688-701.

Williamson, R. C. and Downs T. (1990). "Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds". *International Journal of Approximate Reasoning* 4, 89-158.

VIII Appendix

Appendix 1. Distribution Functions of potential outcomes from available data[21].

Let us start with the definition of the distribution functions of equation (5):

$$\begin{aligned} F_{KY_1}(Ky) &= \Pr[KY_1 \leq Ky] \\ F_{KY_0}(Ky) &= \Pr[KY_0 \leq Ky] \\ F_{KY_0|D}(Ky|1) &= \Pr[KY_0 \leq Ky|D=1] \end{aligned} \qquad (5')$$

Recall that we assumed only one pre-treatment period (denoted by $t=0$) and that $Y_1 = (Y_{10}, Y_{11}, \dots, Y_{1T})'$ e $Y_0 = (Y_{00}, Y_{01}, \dots, Y_{0T})'$ are $Tx1$ vectors that contain the potential outcomes for all periods, whereas $X = (X_1', \dots, X_T')'$ is the matrix that contains the covariates for all periods. The variable KY_d represents the transformation of the potential outcomes that do not depend of the unobserved heterogeneity.

First, we apply the law of iterated expectations to condition everything to covariates at all times.

$$F_{KY_1}(Ky) = \mathbb{E}[\Pr[KY_1 \leq Ky|X]]$$

By assumption (S1) we know that the transformation KY_1 is independent to the treatment variable D conditional on the covariates X, so the distribution function can be written as:

$$F_{KY_1}(Ky) = \mathbb{E}[\Pr[KY_1 \leq Ky|X, D=1]]$$

Using the relation between potential and observed outcomes, $Y_t = DY_{1t} + (1-D)Y_{0t}$ for $t = 0, \dots, T$ and the definition of probability$\Pr[A] = E[I\{A\}]$,

$$F_{KY_1}(Ky) = \mathbb{E}[\Pr[KY \leq Ky|X, D=1]] = \mathbb{E}[\mathbb{E}(DI\{KY \leq Ky\}|X, D=1)]$$

Where $Y = (Y_0, Y_1, \dots, Y_T)'$ is $Tx1$ vector that contains all the observed outcomes at all times. Then, we use the following equation: $\mathbb{E}[S|X] = p(X)\mathbb{E}(S|X, D=1) + (1-p(X))\mathbb{E}(S|X, D=0)$, where S is a random variables. Considering that the propensity score is based on

[21] Adapted from Firpo (2007).

pre-treatment characteristics unaffected by the tratment and that we only have one period before the treatment, the previous expression can be written as a function of $p(X_0)$,

$$F_{KY_1}(Ky) = \mathbb{E}\left[\frac{1}{p(X_0)}\mathbb{E}(DI\{KY \leq Ky\}|X)\right]$$

Finally, using the law of iterated expectations one more time, we reach the desired result

$$F_{KY_1}(Ky) = \mathbb{E}\left[\frac{D}{p(X_0)}I\{KY \leq Ky\}\right]$$

We can obtain analogous results for $F_{KY_0}(Ky)$ and $F_{KY_0|D}(Ky|1)$ using the same procedure.

Appendix 2. Cambanis-Simons-Stout Inequality (CSS)

Lemma: Let R and S two random variables with fixed and known marginal distributions, F_R y F_S; respectivamente. Under the assumption that both R and S have finite variances, it follows that:

$$\int_0^1 F_R^{-1}(1-u)F_S^{-1}(u)du \leq E(RS) \leq \int_0^1 F_R^{-1}(u)F_S^{-1}(u)du$$

These bounds are sharp and finite.

Appendix 3. Application of the CSS inequality to the transformed model[22]

Le tus begin from the results shown in (5). If we denote $W = 1/p(X_0)$ and $V = 1/[1 - p(X_0)]$ and apply the law of iterated expectations:

$$\begin{aligned} F_{KY_1}(Ky) &= E[DE(I\{KY \leq Ky\}W|D)] \\ F_{KY_0}(Ky) &= E[(1-D)E(I\{KY \leq Ky\}V|D)] \\ F_{KY_0|D}(Ky|1) &= \frac{1}{p_1}E[(1-D)E(I\{KY \leq Ky\}V/W|D)] \end{aligned} \qquad (5'')$$

Each expression has the conditional expectation of the product of two random variables ($I\{KY \leq Ky\}$ with W, V y V/W, respectively) so we can apply directly Lemma III to obtain the result shown in Theorem III.1.

It follows that, for $d = 0,1$, $B_{Kd}^I \leq F_{KY_d}(Ky) \leq B_{Kd}^S$, where:

$$B_{K1}^I = E\left[D\int_0^1 F_{I_{KY}|D}^{-1}(1-u|D)F_{W|D}^{-1}(u|D)du\right]$$

$$B_{K1}^S = E\left[D\int_0^1 F_{I_{KY}|D}^{-1}(u|D)F_{W|D}^{-1}(u|D)du\right]$$

$$B_{K0}^I = E\left[(1-D)\int_0^1 F_{I_{KY}|D}^{-1}(1-u|D)F_{V|D}^{-1}(u|D)du\right]$$

$$B_{K0}^S = E\left[(1-D)\int_0^1 F_{I_{KY}|D}^{-1}(u|D)F_{V|D}^{-1}(u|D)du\right]$$

Furthermore, $F_{KY_1|D}(Ky|1) = E[DI_{KY}]/p_1$ is identified, whereas $F_{KY_0|D}(Ky|1)$ is partially identified through $B_{K0|D}^I \leq F_{KY_0|D}(Ky|1) \leq B_{K0|D}^S$, where:

$$B_{K0|D}^I = \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{I_{KY}|D}^{-1}(1-u|D)F_{V/W|D}^{-1}(u|D)du\right]$$

$$B_{K0|D}^S = \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{I_{KY}|D}^{-1}(u|D)F_{V/W|D}^{-1}(u|D)du\right]$$

Where $I_{KY} = I\{KY \leq Ky\}$. To obtain Theorem III.1, it is precise to note that:

$$F_{I_{KY}|D}^{-1}(u|D) = \begin{cases} 0, & \text{para } u \in [0,1 - F_{KY|D}(Ky|D)[\\ 1, & \text{para } u \in [1 - F_{KY|D}(Ky|D), 1] \end{cases}$$

[22] Adapted from Fan *et al.* (2014a).

Finally, replacing this equality in the previous result, we obtain the bounds for $F_{KY_1}(Ky)$, $F_{KY_0}(Ky)$ and $F_{KY_0|D}(Ky|1)$.

Appendix 4. Theorem 3.2 of Fan *et al.* (2014a)

(i) Let $\mu_d(g) \equiv E\left[g\left(\tilde{Y}_d\right)\right]$. So, $\mu_d^L(g) \leq \mu_d(g) \leq \mu_d^U(g)$, for $d = 0,1$, and:

$$\mu_1^L(g) = E\left[D\int_0^1 F_{g(\tilde{Y})|D}^{-1}(1-u|D)F_{W|D}^{-1}(u|D)du\right]$$

$$\mu_1^U(g) = E\left[D\int_0^1 F_{g(\tilde{Y})|D}^{-1}(u|D)F_{W|D}^{-1}(u|D)du\right]$$

$$\mu_0^L(g) = E\left[(1-D)\int_0^1 F_{g(\tilde{Y})|D}^{-1}(1-u|D)F_{V|D}^{-1}(u|D)du\right]$$

$$\mu_0^U(g) = E\left[(1-D)\int_0^1 F_{g(\tilde{Y})|D}^{-1}(u|D)F_{V|D}^{-1}(u|D)du\right]$$

Without additional information, these bounds are sharp.

(ii) Let $\mu_{d|1}(g) \equiv E\left[g\left(\tilde{Y}_d\right)|D=1\right]$. $\mu_{1|1}(g)$ is identified: $\mu_{1|1}(g) = E\left[Dg\left(\tilde{Y}\right)\right]/p_1$ *y* $\mu_{0|1}^L(g) \leq \mu_{0|1}(g) \leq \mu_{0|1}^U(g)$, where:

$$\mu_{0|1}^L(g) = \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{g(\tilde{Y})|D}^{-1}(1-u|D)F_{V/W|D}^{-1}(u|D)du\right]$$

$$\mu_{0|1}^U(g) = \frac{1}{p_1}E\left[(1-D)\int_0^1 F_{g(\tilde{Y})|D}^{-1}(u|D)F_{V/W|D}^{-1}(u|D)du\right]$$

Without additional information, these bounds are sharp.

Appendix 5. Estimators of the bounds of mean counterfactuals

To obtain consistent estimators of the propensity scores we can use the covariate data set. Under the assumption that the treatment participation depends only on pre-treatment characteristics, a consistent estimator, $\hat{p}(x_0)$, can be obtained using the pre-treatment covariate data set[23].

Fan *et al.* (2014b) define, first, the estimated quantile function of the *propensity score* conditional to $D = d$ as $\hat{F}^{-1}_{p(x_0)|D}(u|d) = \inf\{a: \hat{F}_{p(x_0)|D}(a|d) > u\}$ where $\hat{F}_{p(x_0)|D}(a|d)$ is the estimated cumulative distribution function of the propensity score given $D = d$. For $d = 0,1$; this estimated function can be expressed as:

$$\hat{F}_{p(x_0)|D}(a|d) = \frac{\sum_{i=1}^{N}\{\hat{p}(x_0) \leq a, D_i = d\}}{N\hat{p}_d}$$

Using the estimated quantile function of the *propensity score*, Fan *et al.* (2014b) obtained the estimators for the rest of quantile functions as follows:

$$\hat{F}^{-1}_{W|D}(u|d) = \frac{1}{\hat{F}_{p(x_0)|D}(1-u|d)}$$

$$\hat{F}^{-1}_{V|D}(u|d) = \frac{1}{1-\hat{F}_{p(x_0)|D}(u|d)}$$

$$\hat{F}^{-1}_{V/W|D}(u|d) = \frac{\hat{F}_{p(x_0)|D}(u|d)}{1-\hat{F}_{p(x_0)|D}(u|d)}$$

[23] The extension to the case where the treatment participation depends on the "t"th difference of the observed characteristics is straightforward. In this case, we should obtain the *propensity score* based on this differences and, then, we can perform a weighted average (by population) of the *propensity scores.* Finally, we can obtain the quantile functions analogously as Corvalán *et al.* (2015) and Fan *et al.* (2014b).