Bounding Causal Effects in Ecological Inference Problems*

ALEJANDRO CORVALAN, EMERSON MELO, ROBERT SHERMAN AND MATT SHUM

*T**his note illustrates a new method for making causal inferences with ecological data. We show how to combine aggregate outcomes with individual demographics from separate data sources to make causal inferences about individual behavior. In addressing such problems, even under the selection on observables assumption often made in the treatment effects literature, it is not possible to identify causal effects of interest. However, recent results from the partial identification literature provide sharp bounds on these causal effects. We apply these bounds to data from Chilean mayoral elections that straddle a 2012 change in Chilean electoral law from compulsory to voluntary voting. Aggregate voting outcomes are combined with individual demographic information from separate data sources to determine the causal effect of the change in the law on voter turnout. The bounds analysis reveals that voluntary voting decreased expected voter turnout, and that other causal effects are overstated if the bounds analysis is ignored.*

Ecological inference (EI) problems are a class of data combination problems in which aggregate outcome information from one data source is combined with individual demographic information from a separate data source to make inferences about individual outcomes. The objectives of EI include description and prediction of individual behavior, as well as causal inference about individual behavior. King (1997) treats EI problems where the principal objective is description of individual behavior in political science applications. King, Rosen and Tanner (2004) contain articles addressing all three objectives from a number of different fields, including political science, economics, and epidemiology.

This note applies new results from the partial identification literature to make causal inferences with ecological data. Specifically, we use new methods in Fan, Sherman and Shum (2014) to infer the causal effect of a change in Chilean voting law on voter turnout in mayoral elections.

In the absence of the data combination problem, standard results from the treatment effects literature can be used to investigate causal effects. Specifically, if data on outcomes and covariates are observed in the same data set, then it is straightforward, under standard assumptions using known methods, to identify and consistently estimate the usual causal effects of interest, such as the average treatment effect (*ATE*) or the average treatment effect on the treated (*ATT*). For example, one could apply propensity score methods under a standard selection on observables assumption, as in Rosenbaum and Rubin (1983).

* Alejandro Corvalan, Professor of Economics, Department of Economics, Universidad Diego Portales, Av. Santa Clara 797 Huechuraba, Santiago 8580000, Chile (alejandro.corvalan@udp.cl). Emerson Melo, Assistant Professor of Economics, Department of Economics, Indiana University, 307 Wyllie Hall, 100 S Woodlawn, Bloomington, IN 47408, USA (emelo@iu.edu). Robert Sherman, Professor of Economics and Statistics, Division of Humanities and Social Sciences, California Institute of Technology, M/S 228-77, Pasadena, CA 91125, USA (sherman@amdg.caltech. edu). Matt Shum, Professor of Economics, Division of Humanities and Social Sciences, California Institute of Technology, M/S 228-77, Pasadena, CA 91125, USA (mshum@caltech.edu). Corvalan acknowledges financial support from the Institute for Research in Market Imperfections and Public Polity, ICM IS130002.

However, Fan et al. (2014) show that these causal effects, even under the selection on observables assumption, cannot be identified when aggregate outcome data is combined with individual demographic data from separate sources. The information lost through aggregation precludes identification. However, these authors also establish upper and lower bounds on *ATE* and *ATT*, which are valid under data combination. Moreover, these bounds are sharp, meaning that they are the narrowest bounds possible under the maintained assumptions.

We apply these results to our ecological data to estimate bounds on causal effects of the change from compulsory to voluntary voting on turnout in Chilean mayoral elections. In this application, aggregate turnout data must be combined with individual-level census data in order to make causal inferences about the effect of this policy change on voter turnout. For Chile as a whole, the standard difference analysis estimates almost a 27 percent decrease in turnout for the voting-age population under the new law. The robust bounds analysis, on the other hand, estimates anywhere from a 15 percent decrease to 1.2 percent increase in turnout. We show that this pattern holds for many other subsets of the population: ignoring the bounds analysis results in an overstatement of the negative effect of the change from compulsory to voluntary voting on turnout for the voting-age population under the new law.

THE EI FRAMEWORK AND BOUNDING CAUSAL EFFECTS

Here, we introduce the EI model considered in this paper. Then, using the results in Fan et al. (2014), we define sharp population bounds on *ATE* and *ATT* and show how to estimate these bounds.

Let D denote an observed binary treatment assignment indicator. That is, $D = 1$ if an individual is assigned to the treatment group and $D = 0$ if an individual is assigned to the control group. We adopt Rubin's "potential outcomes" approach to describe causal effects. This approach views each individual as having a treatment outcome Y_1 and a control outcome Y_0, but only one of Y_1 and Y_0 is actually observed. The observed outcome is represented as $Y = Y_1 D + Y_0(1 - D)$. Let Z denote observed covariates, which can affect both D and (Y_1, Y_0).

The standard "potential outcomes" approach requires that the analyst observe (Y, D, Z) for each individual in the sample. In the EI context, however, we observe *separate* outcome and covariate data sets. The outcome data set contains (Y, D) while the covariate data set contains (D, Z). Both data sets contain the treatment variable D that links the two sources of information. The objective is to combine these data sources to make inferences about the effect of treatment on outcomes.[1]

Next, we state two assumptions maintained in what follows. Let Z denote the support of the covariate vector Z. For each $z \in Z$, let $p(z) = \mathbb{P}\{D = 1 | Z = z\}$, the so-called propensity score.

A1. SELECTION ON OBSERVABLES: For each $z \in Z$, (Y_1, Y_0) is independent of D given $Z = z$.

A2. OVERLAP: For each $z \in Z$, $0 < p(z) < 1$.

Assumptions A1 and A2 are familiar from the treatment effects literature.

The selection on observables assumption A1 constitutes a "middle ground" between the benchmark of pure randomization and the opposite extreme of pure selection bias. Pure randomization is the assumption that (Y_1, Y_0, Z), the vector of potential outcomes and

---

[1] In some applications, we observe separate outcome and covariate data sets for each treatment. That is, we observe $(Y_1, D = 1)$ and $(D = 1, Z)$, the treatment outcomes and covariates, in separate data sets. Likewise, we observe $(Y_0, D = 0)$ and $(D = 0, Z)$, the control outcomes and covariates, in separate data sets.

observed covariates, is independent of treatment assignment, *D*. This ideal is achieved, for example, in randomized trials or laboratory experiments, and guarantees that the joint distribution of all variables that affect the outcome of interest – observed and unobserved – is the same in both the treatment and control groups. Under pure randomization, no variables are confounded with treatment assignment. This allows the researcher to attribute differences between treatment and control outcomes to the only difference between the two groups, namely, the treatment itself. In other words, under pure randomization, causal effects of the treatment can be inferred from simple comparisons of treatment and control outcomes. On the other hand, under pure selection bias, both observed and unobserved variables are confounded with treatment assignment, making causal effects of the treatment impossible to infer from simple outcome comparisons. By contrast, when A1 holds, there is *conditional* randomization: given *Z*, there are no confounding variables. That is, given *Z*, the joint distribution of all *unobserved* variables that affect outcomes is the same in the treatment and control groups. A1 allows *observed* variables to be confounded with the treatment in the sense that the joint distribution of observed variables is allowed to be different in the treatment and control groups. However, A1 rules out the possibility that given *Z*, unobserved variables make selection into the treatment group more (or less) likely than selection into the control group.

The overlap assumption A2 states that for each $z \in Z$, there is a positive probability that some individual is assigned to the treatment group and a positive probability that some individual is assigned to the control group. Assumption A2 guarantees that in large samples there will be both treatment and control outcomes for each $z \in Z$. Assumptions A1 and A2 make valid comparison of treatment and control outcomes possible for each $z \in Z$.

Fan et al. (2014) show that under data combination, even if A1 and A2 hold, common causal effects like *ATE* and *ATT* cannot be identified. Instead, they derive sharp upper and lower bounds on such causal effects using inequalities from the copula literature. We now develop these bounds.

Recall the propensity score $p(Z) = \mathbb{P}\{D=1 | Z\}$. Let $W = 1/p(Z)$ and $V = 1/[1-p(Z)]$. Define $p_1 = \mathbb{P}\{D=1\}$, the marginal probability of receiving treatment. Define $p_0 = 1 - p_1$. Foreshadowing our application, we develop notation for the special but common case in which the treatment and control outcomes Y_1 and Y_0, and therefore the observed outcomes Y, are binary.

Define $p_{00} = \mathbb{P}\{Y=0 | D=0\}$, $p_{01} = \mathbb{P}\{Y=0 | D=1\}$, and $p_{11} = \mathbb{P}\{Y=1, D=1\}$. Let *X* denote an arbitrary random variable. For $d = 0, 1$, write $F_{X|D}(\cdot | d)$ for the cumulative distribution function of *X* given $D = d$. Write $Q_{X|D}(\cdot | d)$ for the quantile function of *X* given $D = d$. Define the average treatment effect *ATE* and the average treatment effect on the treated *ATT* as follows:

$$ATE \equiv \mathbb{E}\ (Y_1 - Y_0) = \mathbb{P}\{Y_1 = 1\} - \mathbb{P}\{Y_0 = 1\},$$

$$ATT \equiv \mathbb{E}\ (Y_1 - Y_0 \mid D = 1) = \mathbb{P}\{Y_1 = 1 \mid D = 1\} - \mathbb{P}\{Y_0 = 1 | D = 1\}.$$

The following result is a special case of theorem 3.2 in Fan et al. (2014).

THEOREM 1: Suppose $\text{Var}(X) < \infty$ and $\text{Var}(V) < \infty$. If A1 and A2 hold, then

$$\mu_1^L - \mu_0^U \leq ATE \leq \mu_1^U - \mu_0^L,$$

$$p_{11}/p_1 - \mu_{0|1}^U \leq ATT \leq p_{11}/p_1 - \mu_{0|1}^L,$$

where

$$\mu_1^L = p_1 \int_0^{p_{01}} Q_{W|D}(u \mid 1) du,$$

$$\mu_1^U = p_1 \int_{p_{01}}^1 Q_{W|D}(u \mid 1) du,$$

$$\mu_0^L = p_0 \int_0^{p_{00}} Q_{V|D}(u \mid 0) du,$$

$$\mu_0^U = p_0 \int_{p_{00}}^1 Q_{V|D}(u \mid 0) du,$$

$$\mu_{0|1}^L = \frac{p_0}{p_1} \int_0^{p_{00}} Q_{V/W|D}(u \mid 0) du,$$

$$\mu_{0|1}^U = \frac{p_0}{p_1} \int_{p_{00}}^1 Q_{V/W|D}(u \mid 0) du.$$

Let (Y_i, D_i), $i = 1, \ldots, n_1$ denote iid observations of outcome and treatment variables from the outcome data set(s). Let (D_j, Z_j), $j = 1, \ldots, n_2$ denote iid observations of treatment and demographic variables from the covariate data set(s). We estimate the bounds in Theorem 1 by plugging in the sample estimates for the statistical objects in the formulas. Specifically, we use (Y_i, D_i), $i = 1, \ldots, n_1$ to construct the sample proportions $\hat{p}_1, \hat{p}_0, \hat{p}_{01}, \hat{p}_{00},$ and $\hat{p}_{11}$. For example, $\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \{D_i = 1\}$, $\hat{p}_{01} = \frac{1}{n_1 \hat{p}_1} \sum_{i=1}^{n_1} \{Y_i = 0, D_i = 1\}$, $\hat{p}_{11} = \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_i = 1, D_i = 1\}$, and so on.

We use (D_j, Z_j), $j = 1, \ldots, n_2$ to construct $\hat{p}(Z)$, a consistent estimator of the propensity score. There are many ways to estimate the propensity score. One can use parametric estimation procedures like probit or logit, semiparametric estimation procedures, or nonparametric estimation procedures. The estimated quantile functions above are functions of the estimated quantile function of the propensity score. For ease of notation, define $P = p(Z)$. For $d = 0, 1$, we define the estimated quantile function of P given $D = d$ to be $\hat{Q}_{P|D}(u \mid d) = \inf\{a : \hat{F}_{P|D}(a \mid d) > u\}$, where $\hat{F}_{P|D}(\cdot \mid d)$ is the estimated empirical cumulative distribution function of P given $D = d$. That is, with $\hat{P}_j = \hat{p}(Z_j)$, $\hat{F}_{P|D}(a \mid d) = \frac{1}{n_2 \hat{p}_d} \sum_{j=1}^{n_2} \{\hat{P}_j \le a, D_j = d\}$. Using the fact that W is a monotone decreasing function of P, and V and V/W are monotone increasing functions of P, we get that

$$\hat{Q}_{W|D}(u \mid d) = 1/\hat{Q}_{P|D}(1-u \mid d),$$

$$\hat{Q}_{V|D}(u \mid d) = 1/[1-\hat{Q}_{P|D}(u \mid d)],$$

$$\hat{Q}_{V/W|D}(u \mid d) = \hat{Q}_{P|D}(u \mid d)/[1-\hat{Q}_{P|D}(u \mid d)].$$

Finally, the integrals in the expressions above are numerical integrals over the indicated subsets of the unit interval.

Theorems 6.1 and 6.2 in Fan et al. (2016) can be used to prove that the vector of lower and upper bound estimators for both *ATE* and *ATT* are $\sqrt{n}$−consistent and jointly asymptotically normally distributed estimators of their population counterparts. This result permits us to apply the methods in Stoye (2009) to compute asymptotic confidence intervals for *ATE* and *ATT*.

APPLICATION TO CHILEAN VOTING SYSTEM REFORM

From the time democracy was reintroduced in Chile in 1989 until 2012, registration was voluntary while voting was compulsory for registered voters. In early 2012, the law was changed making registration automatic and voting voluntary in presidential, parliamentary, and municipal elections. What effect did the change in the law have on election turnout? We answer this question using the new methods presented in the last section. Since the first elections under the new system were the municipal elections in 2012, we focus our analysis on the most important of the municipal races, namely, the races for mayor.

Chile is divided into 15 regions, which are subdivided into communes or counties. Each commune is governed by a municipality headed by a mayor and a municipal council. Municipal elections in Chile have taken place every four years since 1992. In each election, both the mayor and the council members are elected. Since 2004, the mayor has been elected separately from the council members. Mayoral candidates compete for one seat in each commune and are elected under plurality rule.

We have aggregate voting data for the first mayoral elections under voluntary voting in 2012 as well as aggregate voting data for the first direct mayoral election in 2004, when voting was still compulsory. Our source of voting data is Instituto Nacional de Estadisticas (INE), the Chilean National Statistics Office. Our source of covariate data is Encuesta de Caracterizacion Socioeconomica Nacional (CASEN), the most complete Chilean socioeconomic survey. This survey is conducted by the Chilean government every two to three years in all the communes in the country. Unlike the aggregate voting data from INE, the CASEN data is individual-level data. Corresponding to aggregate voting data in the 2004 election, we use the 2003 CASEN survey, with a sample size of 257,077. Corresponding to aggregate voting data in the 2012 election, we use the 2011 CASEN survey, with a sample size of 200,302. We consider data for individuals who are at least 18 years old, the minimum voting age.

DATA ANALYSIS AND RESULTS

A Naïve Measure

A naïve measure of the causal effect of the new voting law on turnout is the simple difference between turnout proportions in 2012 and 2004. This measure is an unbiased estimate of the causal effect of the change from compulsory to voluntary voting provided the distribution of all observed and unobserved variables affecting turnout is the same in both election years. But this is implausible. For example, the distribution of household income is different in 2004 and 2012, and income is likely to affect turnout. Age may also be a confounding factor. Indeed, the change to voluntary voting was motivated in part by the desire to increase turnout among young voters. Moreover, even if a strong exogeneity condition like A1 holds, standard linear and binary regression methods are impracticable because of the data combination problem. Similar objections can be raised about simple difference-in-differences methods as well as standard linear and binary regression versions of the difference-in-differences techniques.

A New Approach

Given the inadequacy of the naïve difference measure, we turn to our new approach. In the Chilean voting application, $D = 1$ corresponds to being eligible to vote in voluntary Chilean mayoral elections in 2012 and $D = 0$ corresponds to being eligible to vote in compulsory Chilean mayoral elections in 2004. Y_1 is a binary outcome equal to unity if, under voluntary

TABLE 1   *Summary Statistics for Chile*

		2003		2011	
Variables	Description	Mean	SD	Mean	SD
Loginc	Log of annual household income	11.1	1.2	11.9	1.2
Age	In years	42.7	17.3	44.3	18.0
Educ	Completed years of schooling	8.7	4.4	10.1	4.3
Gender	1 If female	0.51	0.49	0.53	0.49
Unemp	1 If unemployed	0.05	0.22	0.04	0.20
Married	1 If married	0.60	0.49	0.55	0.49
Sample size		173,625		144,428	

voting in 2012, an eligible voter turns out to vote, and 0 otherwise. The control outcome Y_0 is a binary outcome equal to unity if, under compulsory voting in 2004, an eligible voter turns out to vote, and 0 otherwise.[2]

Consider assumptions A1 and A2 in the Chilean voting application. Start with A1, the selection on observables assumption. As discussed in the EI Framework and Bounding Causal Effects section, A1 is a weaker assumption than pure randomization. The latter assumption requires that the distribution of all variables that affect turnout, including observed covariates, be the same in the 2004 and 2012 samples. However, Table 1 suggests that the distribution of observed covariates Z is different in the two samples. By contrast, assumption A1 is flexible enough to allow these differences in covariate distributions. At the same time, A1 requires that, conditional on observed covariates like income and age, all unobserved variables that affect turnout are independent of D, the treatment assignment. For example, one unobserved variable that may affect turnout decisions is mayoral candidate quality. Assumption A1 requires that, conditional on observed covariates, the distribution of mayoral candidate quality is the same in 2004 and 2012. While such an assumption is untestable from our data, we feel that by choosing two election cycles relatively close in time (with many mayoral candidates contesting elections in both years), the assumption is plausible with respect to this unobserved variable.

We discuss evidence for overlap assumption A2 after introducing the propensity score model below. *ATE* is the average change in turnout in mayoral elections in 2012 relative to 2004 due to the change from compulsory to voluntary voting, while *ATT* is the average change in turnout in these elections due to the change in voting laws for those eligible to vote in 2012. As the current law makes registration automatic, *ATT* is arguably the more interesting causal measure.

Here, we present estimated bounds on *ATE* and *ATT* for the entire population of Chile as well as for interesting subsets of this population, such as the population of men, the population of women, the 15 regions of Chile, and different age groups.

The observed covariate vector $Z = (Z_1, \ldots, Z_6) = (loginc,\ age,\ educ,\ gender,\ unemp,\ married)$. Table 1 describes each component of Z and gives corresponding summary statistics for Chile.

For a given population subset of interest, let (Y_i, D_i), $i = 1, \ldots, n_1$ denote observations of outcome and treatment variables from the combined INE outcome data sets from 2004 and 2012, and let (D_j, Z_j), $j = 1, \ldots, n_2$ denote observations of treatment and demographic variables

---

[2] For convenience, we use registration as a proxy for voting in 2004. As voting is compulsory in 2004, the differences between those who register and those who vote in 2004 is very small. Also, under compulsory voting, $Y_0 = 0$ if an eligible voter is not registered, as registration is a necessary condition for voting.

TABLE 2   *Estimated Propensity Score Model for Chile*

Variables	Coefficient	SE
Loginc	0.32	0.002
Age	0.008	0.0002
Educ	0.034	0.0006
Gender	0.074	0.005
Unemp	0.12	0.01
Married	−0.15	0.005
n_2	318,053	

from the combined CASEN covariate data sets from 2003 and 2011. Note that n_1 is the sample size of the combined INE outcome data sets and n_2 the sample size of the combined CASEN covariate data sets.

For estimating the bounds on *ATE* and *ATT*, we first estimate the propensity score $p(Z_j) = \mathbb{P}\{D_j = 1 | Z_j\}$ using the CASEN data (D_j, Z_j), $j = 1, \ldots, n_2$ from each population subset of interest. For the country as a whole and for each of the 15 regions of Chile we estimate the propensity score by estimating the probit regression:

$$\mathbb{P}\{D_j = 1 \mid Z_j\} = \Phi(\beta_0 + \beta_1 Z_{1j} + \beta_2 Z_{2j} + \beta_3 Z_{3j} + \beta_4 Z_{4j} + \beta_5 Z_{5j} + \beta_6 Z_{6j}).$$

Note that the probit assumption implies that $\mathbb{P}\{D_j = 1 \mid Z_j\} = \mathbb{P}\{D_j = 1 \mid Z_j'\beta\}$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_6)$. Thus, conditioning on the vector Z_j is equivalent to conditioning on the linear index $Z_j'\beta$. We also estimate propensity score models for men and women separately, and for separate age categories $18-24$, $25-29$, $30-34$, $35-39$, $40-44$, $45-49$, $50-54$, $55-59$, $60-64$, $65-69$, and $70-74$.[3]

Table 2 presents coefficient estimates and standard errors for the probit regression for the entire country. All the variables make a statistically significant marginal contribution and so have power to predict treatment assignment. The first five variables make positive contributions, whereas the married variable makes a negative contribution. The significant positive coefficient on *loginc* implies that *ceteris paribus*, the higher a person's income, the more likely that person is an eligible voter in 2012 rather than in 2004. Similar interpretations can be made for the other variables in the model.

At this point it is convenient to return to the overlap assumption A2. A2 says that for each possible value of Z, there is a positive probability of obtaining an observation from 2012 as well as from 2004. Figure 1 is a plot of the estimated propensity score $\hat{\mathbb{P}}\{D = 1 \mid Z'\hat{\beta}\}$ versus its estimated index $Z'\hat{\beta}$, where $\hat{\beta}$ is the probit maximum likelihood estimator (of β) whose estimated components are given in Table 2. For each possible value of the estimated index $Z'\hat{\beta}$, the estimated propensity score is strictly between 0 and 1, an informal visual confirmation of A2. In Figure 2, we provide corresponding histograms of the estimated propensity scores in both the 2004 and 2012 subsamples. Note that they share large regions of common support, providing additional evidence of the plausibility of A2.

Before discussing the results of the bounds analysis, it is instructive to make the following observation. Suppose that the set of observed covariates Z has no power to predict treatment assignment D. This holds, for example, if D is independent of Z. Now, if D is independent of Z and A1 holds, then (Y_1, Y_0, Z) is independent of D, which is the pure randomization assumption. Under pure randomization, *ATE* and *ATT* are equal and the simple difference estimator

---

[3] The separate models for men and women have the same form as the probit regression above, except that the gender variable Z_4 is dropped from the model. Similarly, the probit regression for the separate age categories has the same form except that the age variable Z_2 is dropped.
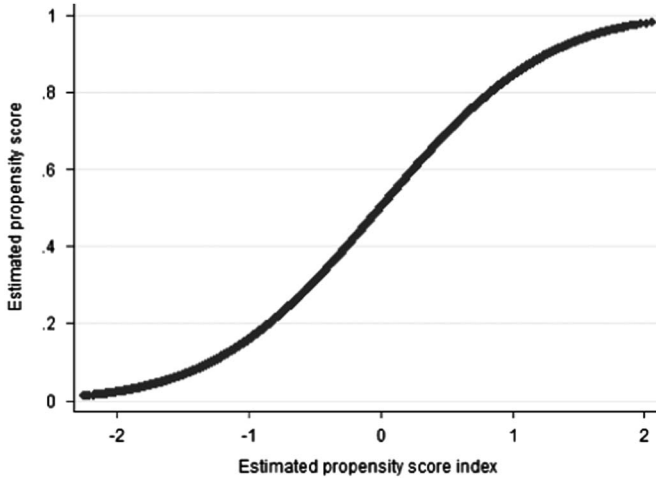
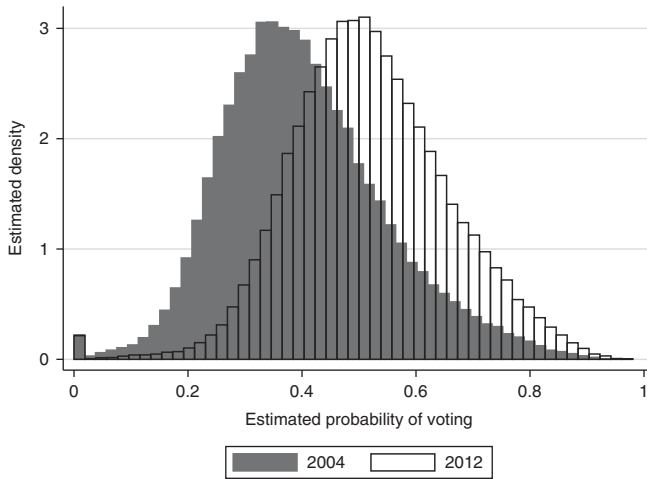Fig. 1. Informal visual confirmation of the overlap assumption A2 for the country as a whole

Fig. 2. Estimated densities for the probability of voting in 2004 and 2012

$$\frac{\sum_{i=1}^{n_1} Y_i D_i}{\sum_{i=1}^{n_1} D_i} - \frac{\sum_{i=1}^{n_1} Y_i(1-D_i)}{\sum_{i=1}^{n_1}(1-D_i)} \text{ is a consistent estimator of both } ATE \text{ and } ATT. \text{ In other words, under pure}$$

randomization, no bounds analysis is needed and estimates of *ATE* and *ATT* are identical. As we now show, our results belie the pure randomization assumption and support the need for bounds analysis.

Figure 3 displays 95 percent confidence intervals for *ATE* and *ATT* for the country as a whole, as well as for men and women separately. For better visual effect, these confidence intervals are represented as boxes, where the length of a box is the length of the corresponding confidence interval. Focus on the box for *ATE* for the country. The ordinate of any point on the top (bottom) of this box is the upper (lower) bound estimate for *ATE* given in Theorem 1
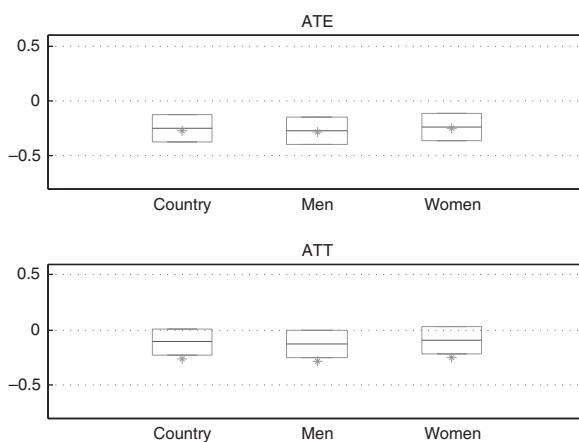
Fig. 3. *95 percent confidence intervals, represented as boxes, for average treatment effect (ATE) and average treatment effect on the treated (ATT) of the change from compulsory to voluntary voting on turnout in Chile, for the country as a whole and by gender*

plus a standard error correction computed using the procedure of Stoye (2009).[4] The box is split in the middle by a line. The starred point represents the simple difference estimate defined in the last paragraph.[5] Corresponding statements apply to the other *ATE* boxes and to the *ATT* boxes.

Consider the *ATE* boxes in Figure 3. For the country as a whole as well as for men and women separately, the simple difference estimates suggest that voluntary voting decreased voter turnout. This suggestion is confirmed by the robust bounds analysis: each 95 percent confidence interval upper bound is below the zero level.

Next consider the *ATT* boxes in Figure 3 and recall that *ATT* may be the more relevant causal measure since registration and therefore eligibility is automatic under current Chilean law. The robust bounds do not contain the simple difference estimates. Under A1, this is strong evidence against the pure randomization assumption and strong evidence for the need for this type of bounds analysis. If the bounds analysis were ignored, the negative effect of voluntary voting on turnout for eligible voters in 2012 would be overstated. In fact, all three *ATT* boxes contain the point 0, although just barely. This suggests that we cannot reject the hypothesis that voluntary voting had no effect on turnout in 2012 at the 5 percent level.

Figure 4 displays bounds results for *ATE* and *ATT* for the 15 regions comprising Chile. As a reference point, the last box in Figure 4 represents the results in Figure 3 for the country as a whole. The results for the individual regions are qualitatively the same as those for the country as a whole. All the *ATE* boxes are below the zero level and, with the exception of Region 15,

---

[4] As mentioned in the EI Framework and Bounding Causal Effects section, the procedure of Stoye (2009) is valid under joint asymptotic normality of the lower and upper bound estimators given in Theorem 1. The joint asymptotic normality results are given in theorem 6.1 and theorem 6.2, respectively, in Fan et al. (2016). The asymptotic standard errors in these theorems are estimated with the bootstrap to produce our standard error corrections. The standard error corrections in this application are typically negligible compared with the length of the bounds.

[5] The turnout difference estimates can be taken as exact population differences. The reason is that they are based on very large sample sizes, making the length of the corresponding confidence intervals 0 for all practical purposes.
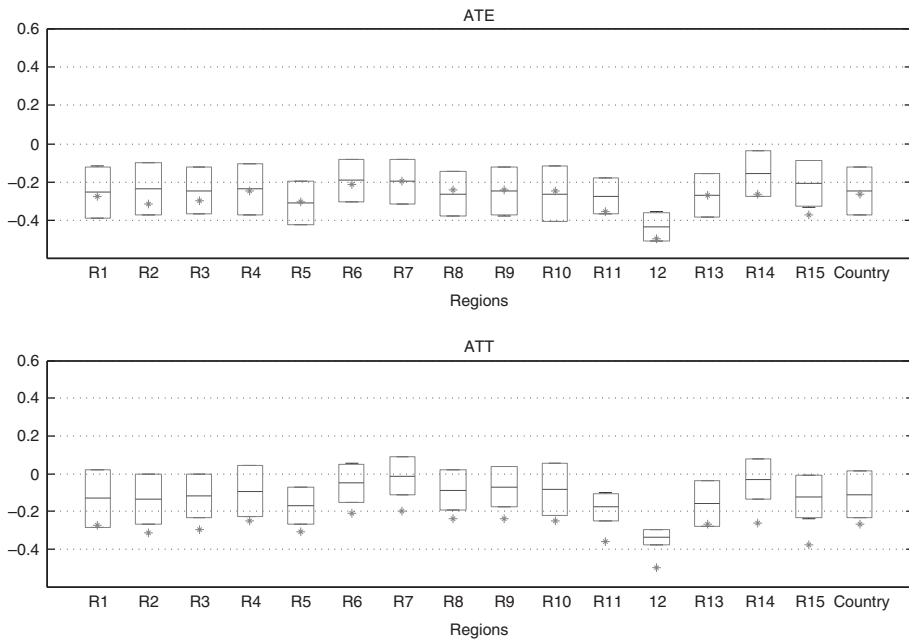
Fig. 4. 95 percent confidence intervals, represented as boxes, for average treatment effect (ATE) and average treatment effect on the treated (ATT) of the change from compulsory to voluntary voting on turnout in Chile, for the country as a whole and by region
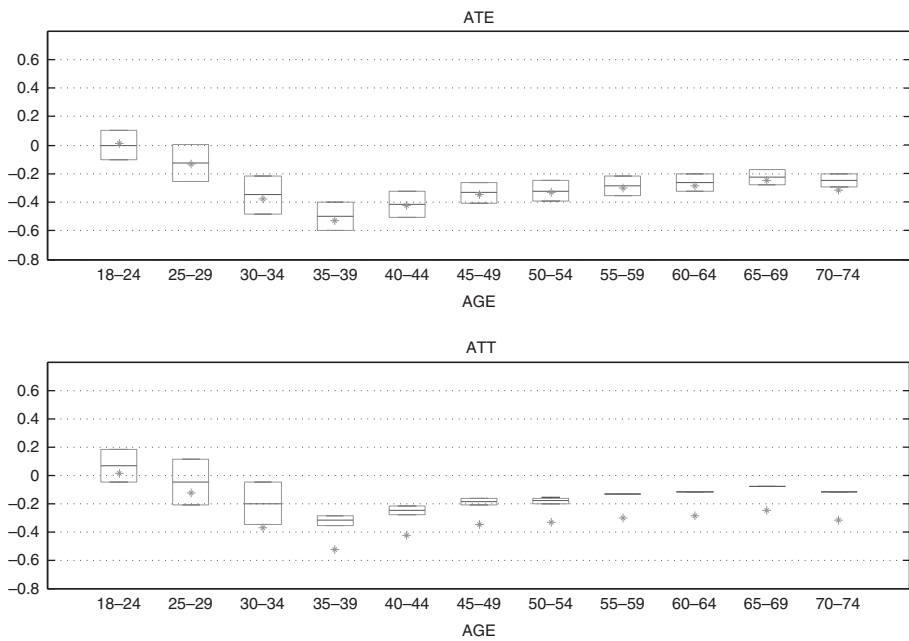

Fig. 5. 95 percent confidence intervals, represented as boxes, for average treatment effect (ATE) and average treatment effect on the treated (ATT) of the change from compulsory to voluntary voting in Chile, by age category

contain the corresponding simple differences estimates. The *ATT* boxes for Regions 5, 11, 12, and 13 are all below the zero level, implying that voluntary voting has, at the 5 percent level, a statistically significant negative effect on turnout in these regions. In all regions except possibly Regions 1 and 13, ignoring the bounds analysis and taking the simple difference estimates at face value overstates the negative effect of voluntary voting on turnout for those eligible to vote in 2012.

Finally, consider Figure 5, which displays results for *ATE* and *ATT* conditional on age. The results are qualitatively similar to those presented in the previous figures. However, focus on the two youngest age categories, and recall that one of the motivations for changing from compulsory to voluntary voting was to try to increase turnout among young voters. The *ATE* and the *ATT* boxes both straddle the zero level for the 18–24 and 25–29 age categories, and the *ATT* box for the 18–24 age category is nearly above the zero level. While not conclusive at the 5 percent significance level, the results do not rule out the possibility that voluntary voting had a positive effect on turnout among younger voters, in line with the intended goals of the policy change.

CONCLUSION

This note shows how to apply new partial identification results from the treatment effects literature on data combination to make inferences about causal effects in EI problems. More broadly, the need to combine different data sources in causal effect modeling appears commonplace in political science. Other potential applications include measuring the effect of introducing electronic voting on vote outcomes, the effects of war on health outcomes, or the effects of political turmoil on economic activity. In all these cases, one needs to combine aggregate (precinct-level, regional-level, or country-level) outcome data with demographic confounders measured at the individual level.

We apply our methodology to bound causal effects of a change from compulsory to voluntary voting on turnout in recent Chilean mayoral elections. Our bounds analysis reveals that the change had a negative effect on expected turnout and that ignoring this analysis leads to overstating the negative effect of the change on those who are eligible to vote under the current voluntary voting laws. We plan to extend our analysis to recent parliamentary and presidential elections in Chile.

REFERENCES

Fan, Yanqin, Robert Sherman, and Matt Shum. 2014. 'Identifying Treatment Effects Under Data Combination'. *Econometrica* 82:811–22.

Fan, Yanqin, Robert Sherman, and Matt Shum. 2016. 'Estimation and Inference in an Ecological Inference Model'. *Journal of Econometric Methods* 5:17–48.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

King, Gary, Ori Rosen, and Martin Tanner. 2004. *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

Rosenbaum, Paul, and Donald Rubin. 1983. 'The Central Role of the Propensity Score in Observational Studies of Causal Effects'. *Biometrica* 70:41–55.

Stoye, Jorg. 2009. 'More on Confidence Intervals for Partially Identified Parameters'. *Econometrica* 77: 1299–315.