How Do Non-experimental Methods of Estimating the Average Treatment Effect Compare to Experimental Methods?

Thipbadee Tantivilaisin

10 May 2019

Abstract

This paper evaluates the validity of the causal estimated average treatment effect using correction on observables (COO) with a non-experimental framework by comparing it to the local average treatment effect (LATE) derived from a randomized control trial using the instrumental variables (IV) method. The data for both these methods were obtained from a subset of a voter mobilization experiment conducted prior to the 2002 U.S. midterm election by Arcenaux et al. The first step of our analysis began with creating a balance table to verify that certain conditions are met before we began our methods. The table verified that the randomization of the experiment was successful for the RCT, but not for the COO due to how the treatment and control groups were defined. The second method used estimated the local average treatment effect using a two-stage least squares regression for the experimental data and a regression for the non-experimental data. Both methods resulted in the average treatment effect being statistically significant in our framework, however, COO overestimated the average treatment effect by 2.10 times the amount of our IV estimate due to selection and omitted variable bias.

1 Introduction

Econometricians have historically used non-experimental methods (such as correction on observables) to try to replicate the results from experimental methods. The flaw of correction on observables is that if we are unable to account for biases that occur naturally compared to a well randomized experiment, then we will have a biased estimate. The obvious question is, exactly how accurate are these non-experimental methods when compared to experimental ones? We are motivated by the fact that non-experimental methods incur a fraction of the cost, have no ethical concerns and are free from mistakes during the experimental process.

This question is answered through the framework of a study administered prior to the 2002 U.S. midterm election conducted by Arcenaux et al. The researchers ran an experiment measuring the effectiveness of being assigned a phone call that encouraged an individual to vote on whether that individual voted in the 2002 U.S. midterm election (we defined this as Assigned Call). They also treated the same experimental data set as non-experimental when they defined individuals picking up the call, listening to the whole message,

and answering the question as the treatment (we defined this as Contact). We defined our experimental and non-experimental data to be the same as what Arcenaux et al. did. Ultimately, the question of whether we can estimate the average treatment effect with non-experimental data will be answered by whether we can estimate the causal average treatment effect of Contact on if the individual voted in 2002.

We begin by creating a balance table for both the experimental and non-experimental data sets. This is done to evaluate the differences between the control and treatment groups for each data set in terms of observable characteristics. With regards to the experimental data which was conducted using a randomized control trial, we can evaluate how well the randomization was organized. This can also be done with the non-experimental data; however, it can be expected that the differences in the observable characteristics are statistically significant. This is inherent due to the way the treatment was defined by the researchers which ultimately led to selection bias. Next we estimated the treatment effect through a OLS regression and a two-stage least squares regression for the experimental data set in addition to a OLS regression for the non-experimental data set. By adding additional covariates, we can see if omitted variable bias is present if the respective coefficients begin to show an obvious pattern of convergence with each additional covariate added. We then compare the estimated local average treatment effect derived from the randomized control trial using the instrumental variables method in contrast to the estimated average treatment effect derived utilizing correction on observables.

When we created the balance tables for the experimental data it was immediately clear that the randomization for this experiment was done well. When it came to the balance table of the non-experimental data, the differences in the sample characteristics of the treatment and control were statistically significant to more than a 99.9% degree of confidence. In order to get the estimated intention to treat effect we first ran a regression of whether the individual voted in 2002 on Assigned Call. Next, to determine the local average treatment effect we ran a two-stage least squares regression where the first stage is the effect of being assigned a call on if the individual picked up and listened to the message. The second stage is the effect of picking up the call and listening to the message on whether that individual voted in the 2002 U.S. midterm election. The estimated intention to treat effect coefficient and estimated local average treatment effect did not change significantly implying that it was not heavily subjected to biases. However, when we added additional covariates to non-experimental data set, however, the estimated average treatment effect changed significantly. We then compare the estimated local average treatment and estimated treatment effect coefficients with a significance test.

2 Data

The data used in this paper comes from a paper published in 2006 titled, "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment" by Arcenaux et al. The intent of this study was to analyze the effectiveness of encouraging individuals to vote through phone calls. The

location of this experiment was conducted in the states of Iowa and Michigan before the 2002 U.S. midterm elections. Households that contained one or two registered voters were randomly selected into the treatment and control groups. When there were two voters in a household, only one of them were assigned to a group and the other was ignored to simplify statistical analysis. Members of the treatment group received a call with a non partisan "get-out-the-vote" message:

> Hello, may I speak with (name of person) please? Hi. This is (caller's name) calling from Vote 2002, a nonpartisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday? (Arcenaux et al., 2006, p. 41)

Whether an individual was in the treatment or control group is denoted by the variable called Assigned Call. Individuals were identified as contacted (embodied by the variable Contact) if they listened to the whole message and responded to the question. Therefore, the non-experimental data is a subset of the treatment group in the experimental data set. The members in the control group were simply not called and essentially not a part of the non-experimental data set.

Once the election ended, the researchers gathered the observable characteristics from voter registration lists on the individuals selected into the control and treatment groups. They were able to extract information about past voting history, general voting characteristics and location. The justification for the non-inclusion of the usual estimators for voter turnout is that a voter's past voting history cancels out the absence of usual estimators (Plutzer 2002).

3 Methods

Assuming randomization of the experimental data set was done correctly, we can estimate the average treatment effect for the non-experimental and intent to treat effect for the experimental data through regression. We begin by running the following:

$$Vote02_i = \beta_0 + \beta_1 treatment_i + \beta^{'} X + u_i \tag{1}$$

Equation 1 represents the general form of the regression that was ran for both the experimental and non-experimental data sets. The treatment variable can be substituted in for Assigned Call when it is the experimental data, and Contact when it is the non-experimental. Let β_1 be the effect of respective treatment effects holding all else constant. Let X be a set of covariates that include age, gender and past voting behavior.

We hypothesize that if the randomization was done correctly for the experimental data, the only statistically significant differences between the two groups will be that the treatment group received the call and the control group did not. What this tells us is that if we were to add additional covariates into our regression where Assigned Call is our β_1, then it should remain consistent.

However, Assigned Call does not give us a direct comparison to Contact as Assigned Call is the estimated intent to treat effect while Contact is the estimated average treatment effect. Thus, we run the two-stage least squares regression in order to derive the instrumental variable estimator or estimated local average treatment effect which results in a fairer comparison. The first and second stages are the following:

$$\hat{Contact}_i = \phi_0 + \phi_1 AssignedCall \tag{2}$$

$$Vote02_i = \pi_0 + \pi_1 \hat{Contact}_i + \pi^{'} X + \lambda_i \tag{3}$$

Equation 2 represents the first stage and tells us the estimated effect of assigning an individual a call on whether they picked up and listened to the whole message. The second stage is the effect of picking up the call and listening to the message on whether that individual voted in the 2002 U.S. midterm election. Once again, we let X represent the set of covariates that includes age, gender and past voting history. We derive the estimated local average treatment effect by substituting the first stage equation into the second stage. Thus, our estimated local average treatment effect is embodied in the second stage with the Contact coefficient.

When we add additional covariates to equations 1 and 2, the change in the coefficients with regards to each additional covariate should dictate how much bias is in our model. We expect the estimated treatment effects derived from the randomized control trial to not change significantly. However, we do expect the estimated average treatment effect derived from the non-experimental data set to change significantly due to the inherent presence of bias in non-experimental methods. Once we account for these biases by adding additional covariates, we can compare the estimated effects.

Finally, we can answer the question of whether we can estimate the causal effect of treatment using non-experimental data. By comparing π_1 which represents the estimated local average treatment effect derived from the randomized control trial using a two-stage least squares regression and β_1 representing the average treatment effect of Contact after its biases has been accounted for by adding covariates. Furthermore, π_1 and β_1 should reach the same conclusion when we run a hypothesis test where $H_0 : \pi_1 = 0$ and $H_0 : \beta_1 = 0$. However, if we fail to reject the null for both these tests, then it is inconclusive whether or not we have reached the same conclusion. Therefore, we must run another test where $H_0 : \beta_1 = \pi_1$ to answer our question.

4 Results

In order for the randomized control trial to give a good estimate of the treatment effect, randomization between the control and treatment groups must have been done correctly. The significance of which, being that the only differences between the two groups are that one was treated and the other was not, therefore it should not be biased in any way. Table 1 presents the observable characteristics categorized by the control group and treatment group in the experimental data set. In addition, the table also displays the differences between the two groups and their p-values such that they are equal for each covariate. We can observe that Age, Newly Registered and Voted in 2000 are well beyond the standard 95% confidence level. However, we also observe that Female and Voted in 1998 are 0.0370 and 0.0345 respectively. With these values we would reject H_0 at a 95% confidence level, but fail to reject at the 97% level which is certainly good enough as this could have resulted due to random chance from a bad drawing of the overall sample. This would prove to be true, as the original balancing from Arcenaux et al. was done correctly.

This implies that we met the basic criteria that satisfies all of the benefits that the randomized control trial brings. The first of which being that the sample size is large enough ($n_c = 85628$, $n_t = 14990$) and the randomization was done correctly (with ~97% level of confidence). What we ultimately want to measure is how a treatment, which in this framework is being assigned a call and its effect on whether the individual voted. However, the fundamental problem of causal inference is that we are unable to obtain both $\{Y_i \,|Z_i = 1\}$ and $\{Y_i \,|Z_i = 0\}$ for any same individual such that $Z_i = 1$ is being assigned to the treatment group. However, with a randomized control trial we are able to estimate the average treatment effect. This same treatment effect is present in section 3 equation 1 with β_1 being the average treatment effect of Assigned Call.

Since the sample size is large and the randomization was properly performed, we can expect that β_1 is very close to the true treatment effect of being assigned a call. Thus, it is not subject to any biases such as omitted variable bias. Therefore, when we add additional covariates to equation 1, we should expect that β_1 to not vary much. We can observe this fact in Table 2, where β_1 only varies within the range of 1.73 percentage points and 1.24 percentage points. Next, we test whether β_1 for Assigned Call is a statistically significant coefficient by testing $H_0 : \beta_1 = 0$. When all the covariates have been added to this regression, we get $p < 0.01$, therefore rejecting H_0 and ultimately suggesting that being assigned to get a call has a statistically significant impact on voting in 2002.

However, as we mentioned earlier Assigned Call is only the estimated intention to treat, thus not everyone that was assigned into the treatment group in the experimental data set received the treatment so we can not make the direct comparison between it and the average treatment effect derived from correction on observables. In fact, only 46.26% of individuals received the treatment, therefore we know for a fact that the estimate has attenuation bias. In order for us to answer our question, we have to get the instrumental variable estimator by running a two-stage least squares regression.

We ran the two-stage least squares regression to derive the instrumental variable estimator with the

assumption that it will eliminate the attenuation bias from the estimated intention to treat giving us the estimated local average treatment effect. We observe this in Table 5, and see that the estimated local average treatment effect is 2.13 times greater after having added all the covariates. We also observe that the estimated local average treatment effect ranges from 3.74 to 2.65 percentage points which is not too significant of a change. We then tested whether the estimated local average treatment effect coefficient π_1 is statically significant or not, which it is as we get a p-value less than 0.1.

Moving onto the analysis of the non-experimental data, we went on to evaluate it in terms of how statistically significant the differences in the observational characteristics are. The format of the Table 3 is the same as Table 1, where the means of each covariate are categorized by control and treatment in addition to a column for the differences between them. Once again, it displays the p-value such that they are equal for each observable characteristics. We immediately observed that all of the differences in the observable characteristics between the treatment and control groups are statistically significant to more than a 99.9% level of certainty.

As mentioned earlier in this paper, this was expected due to how the individuals had the ability to choose whether they wanted to pick up the call or not. This choice had thereby caused the distribution of individuals into the treatment and control groups to not be randomized. The side effect of having the selection bias is that there is also omitted variable bias when we run our regression model. This means that the more covariates we are able to add into the regression, the closer the β_1 treatment coefficient should be to its true value, which at this point we assumed to be close to the estimated local average treatment effect π_1.

We run the regression and see that Contact's effect is 5 times greater than that of Assigned Call's effect. Once we add more covariates we can see an immediate trend that β_1 decreases significantly. The β_1 coefficient of Contact went from a 11.7 percentage points effect on voting in 2002 to 5.59 percentage points. So we find out that there is quite a large amount of omitted variable bias that occurs in our model. Mathematically, the way that omitted variable bias affects estimated treatment effect is embodied by the following equation: $\hat{\beta}_1 = \beta_1^* + \beta_x \alpha_i$. We let $\hat{\beta}_1$ be the biased effect of treatment, let β_1^* be the true effect of the treatment effect and let α_i be difference between the control and treatment groups for each covariate. The amount that $\hat{\beta}_1$ deviates away from β_1^* is the effect of the omitted variable on voting in 2002 $[\beta_x \,|x \epsilon K]$ multiplied by α_i such that K is the set of all omitted variables.

An example of an omitted variable that created a large overestimation in our estimates would be age. Intuitively, we know that the older an individual gets, the more likely that they are to vote. Hypothetically if we ran the regression of $vote02_i = \beta_0 + \beta_{age} age_i + u_i$ such that the sample size was large enough and randomization was done perfectly, we can find the value of β_{age}. We can test our intuition by running that regression with our RCT data, where we know the randomization was done correctly. The value of β_{age} in that regression is 0.0077 with the t-stat of 125.62 so we reject $H_0 : \beta_{age} = 0$ easily and know that our intuition was correct. The next variable that we need is $avg_n [\alpha_i]$ which can be found by looking at table 3. This gives us the value of 5.381. Once multiplied out, we can obtain the value that has biased β_1 where the

treatment is contact from its true value.

The estimated treatment effect of being assigned a call had such little omitted variable bias because there was such a small α_i due to the randomization being done correctly. This effectively made it so that $[\beta_x \, | x \epsilon K]$ was being multiplied by a number very close to zero. In an experiment which was perfectly randomized and $\lim_{n \longrightarrow \infty}$ such that n is the sample size, we can expect no omitted variable bias present in our estimate. This ultimately gives us the true value of the treatment effect and the reason why in our first regression β_1 still changed when we added additional covariates.

Finally, after having verified that all the condition have been met allowing us to use our methods mentioned in section 3, we can compare the estimated local average treatment effect we derived from the experimental data set and the estimated average treatment effect derived from the non-experimental data set. However, we must also acknowledge that while this is the most fair comparison we can make, the effect from the experimental data is only the average treatment effect for the subset of the population that are compliers. The treatment effect we derived from the non-experimental data is estimated treatment effect for the whole population. With this in mind, we observe the estimated local average treatment effect and the estimated average treatment effect to both be statistically significant. Within the context of the framework, this meant that the effect of being assigned a call on compliers and the effect of Contact on whether an individual voted in 2002 is statistically significant. While this told us that the non-experimental method would've reached the same conclusion as experimental method, we must also see the accuracy of the non-experimental method. We observe from Table 4 and Table 5 that the estimated average treatment effect is 2.10 times greater than the estimated local average treatment effect, therefore immediately failing the two mean significance test. Therefore we conclude that while being assigned a call, listening to the whole message and answering the question had a statistically significant effect on voting in the 2002 U.S. midterm election with both the experimental and non-experimental methods, it would seem that the non-experimental method still overestimated the effect by a statistically significant amount.

We must also acknowledge that the randomization also almost failed for the Female and Voted in 1998 covariate, so our experimental data is a bit more biased than we would have liked. Therefore the result of whether the voter mobilization effort actually had a statistically significant impact on an individual's voting outcome is inconclusive. However, we can conclude that we could not get the causal estimated average treatment effect from non-experimental data that was anywhere near as good as using experimental data.

5 Conclusion

We found that the randomization of the experiment was done correctly giving us all the benefits that a well conducted experiment bestows. We were able to retrieve the estimated effect of Assigned Call but what we ultimately wanted was the estimated local average treatment effect of being assigned a call and responding to it which we derived through a two-stage least squares regression. When we added covariates to

the second stage, we found that the estimated effect did not change significantly. When we ran the regression for the non-experimental data and added covariates we saw that the average effect of Contact changed a significant amount. This implied the existence of omitted variable bias present in our model. We reached the conclusion that both these estimates are statistically significant in terms of how they affected voting in the 2002 midterm election. When we finally compared the estimated local average treatment and the estimated average treatment effect, we found that correction on observables had overestimated the effect of Contact by more two times the amount of our benchmark.

Overall, it would prove to essentially be impossible to consistently obtain a causal estimate of the treatment effect using correction on observables with non-experimental data. This is due to the fact that the randomized control trial works because of randomization, which results in the difference of observable characteristics to effectively equal zero. Without randomization, unless there is a scenario in which there are no omitted variables then there will always be bias present in the estimated treatment effect that needs to be accounted for. This is because there are an infinite amount of observable characteristics that affects a person's decision and just as much if not more for non-observable characteristics that we can never account for.

The motivation behind our question was that we preferred to use correction on observables because it costs much less than the randomized control trial. While the randomized control trial is limited due to the costs and ethical issues that may arise (which can hypothetically be overcome), correction on observables is limited due to the fact that it can not produce a causal estimate of the treatment effect due to the absolute impossibility of researchers ever being able to account for every covariate that contributes to the outcome variable.

If the question is of enough importance and interest, it is to the best interest of the researchers and its stakeholders that it be answered not through correction on observables, as it does not produce a causal estimate of the treatment effect. However, if a non-experimental method must be used, the conclusions that one draws from it should be taken with caution.

References

1. Arceneaux, Kevin, Alan Gerber, and Donald Green. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis*, 2006, 14, 37-62

2. Plutzer, E. (2002). Becoming a Habitual Voter: Inertia, Resources, and Growth in Young Adulthood. *American Political Science Review*, 96(1), 41-56

Table 1: Balance Check between groups: Being Assigned Call vs. Not Being Assigned Call

	Control	Treatment	Difference	
	mean	mean	b	p
Age	55.768	56.056	-0.288	0.0844
Female	0.564	0.554	0.009*	0.037
Newly Registered	0.049	0.052	-0.003	0.148
Voted in 1998	0.572	0.581	-0.009*	0.0345
Voted in 2000	0.736	0.739	-0.003	0.4398
Observations	85628	14990	100618	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: The Effect of Being Assigned a Call on on Whether an Individual Voted in the 2002 U.S. Midterm election

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	vote02	vote02	vote02	vote02	vote02	vote02
Assigned Call	0.0173***	0.0156***	0.0150***	0.0157***	0.0124***	0.0124***
	(0.00434)	(0.00423)	(0.00424)	(0.00422)	(0.00386)	(0.00365)
Age		0.00579***	0.00623***	0.00574***	0.00205***	0.00173***
		(7.95e-05)	(8.07e-05)	(8.19e-05)	(7.97e-05)	(7.53e-05)
Female			-0.0261***	-0.0259***	-0.0198***	-0.0184***
			(0.00306)	(0.00304)	(0.00279)	(0.00263)
Newly Registered				-0.217***	-0.0410***	0.167***
				(0.00712)	(0.00665)	(0.00656)
Voted in 1998					0.419***	0.276***
					(0.00306)	(0.00317)
Voted in 2000						0.398***
						(0.00364)
Constant	0.595***	0.273***	0.279***	0.316***	0.269***	0.0580***
	(0.00168)	(0.00473)	(0.00493)	(0.00506)	(0.00465)	(0.00479)
Observations	100,618	100,618	98,075	98,075	98,075	98,075
R-squared	0.000	0.050	0.058	0.066	0.216	0.302

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Balance Check between groups: Picking up the Call vs. Not Picking up the Call

	Control	Treatment	Difference	
	mean	mean	b	p
Age	53.567	58.948	-5.381***	0
Female	0.539	0.572	-0.032***	0.0001
Newly Registered	0.057	0.045	0.012***	0.0007
Voted in 2000	0.703	0.781	-0.078***	0
Voted in 1998	0.546	0.622	-0.076***	0
Observations	8055	6935	14990	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: The Effect of Contact on Whether an Individual Voted in the 2002 U.S. Midterm election

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	vote02	vote02	vote02	vote02	vote02	vote02
Contact	0.117***	0.0894***	0.0739***	0.0746***	0.0666***	0.0559***
	(0.00792)	(0.00785)	(0.00787)	(0.00781)	(0.00723)	(0.00683)
Age		0.00507***	0.00558***	0.00490***	0.00142***	0.00114***
		(0.000208)	(0.000211)	(0.000214)	(0.000211)	(0.000199)
Female			-0.0330***	-0.0325***	-0.0228***	-0.0210***
			(0.00786)	(0.00780)	(0.00722)	(0.00681)
Newly Registered				-0.268***	-0.101***	0.115***
				(0.0180)	(0.0170)	(0.0168)
Voted in 1998					0.393***	0.256***
					(0.00799)	(0.00818)
Voted in 2000						0.401***
						(0.00940)
Constant	0.559***	0.287***	0.299***	0.350***	0.305***	0.0898***
	(0.00539)	(0.0123)	(0.0128)	(0.0131)	(0.0122)	(0.0126)
Observations	14,990	14,990	14,626	14,626	14,626	14,626
R-squared	0.014	0.052	0.057	0.071	0.203	0.292

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5: The Local Average Treatment Effect of Being Assigned a Call on Whether an Individual Voted in the 2002 U.S. Midterm Election

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	vote02	vote02	vote02	vote02	vote02	vote02
Contact	0.0374***	0.0338***	0.0320***	0.0334***	0.0264***	0.0265***
	(0.00938)	(0.00914)	(0.00904)	(0.00900)	(0.00824)	(0.00778)
Age		0.00577***	0.00621***	0.00572***	0.00204***	0.00172***
		(7.97e-05)	(8.08e-05)	(8.21e-05)	(7.98e-05)	(7.54e-05)
Female			-0.0262***	-0.0260***	-0.0199***	-0.0185***
			(0.00306)	(0.00304)	(0.00279)	(0.00263)
Newly Registered				-0.217***	-0.0411***	0.167***
				(0.00712)	(0.00664)	(0.00656)
Voted in 1998					0.419***	0.276***
					(0.00306)	(0.00317)
Voted in 2000						0.398***
						(0.00364)
Constant	0.595***	0.274***	0.280***	0.317***	0.270***	0.0590***
	(0.00167)	(0.00470)	(0.00490)	(0.00503)	(0.00462)	(0.00477)
Observations	100,618	100,618	98,075	98,075	98,075	98,075
R-squared	0.001	0.051	0.058	0.067	0.217	0.302

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1