Benefit Analysis of Regression Methods and Instrumental Variables

By: Soham Shewale

January 31, 2019

Abstract

Recently, there has been a push to get more people to vote in elections, but how effective is this crusade to get people to go to the polls. This paper looks at a study done by Arceneaux, Kevin, Alan Gerber, and Donald Green called “Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment”. In this data, the experiment is to see the effect of two treatments, getting the call and listening to the entire call, on the outcome of voting in the 2002 election. Within the regression with these two variables, people who were assigned to get the call are more likely to vote in the 2002 election, but people who listened to the phone call are even more likely to vote. The variable of listening to the call all the way through is a representation of a subpopulation because of the selection bias of people selecting themselves into the treatment. People select themselves into the treatment because they are different than those assigned to get a call, this is due to an omitted variable. Another major point of this study is to see if the instrumental variable is a good way to swerve around the problem of selection and omitted variable bias.

Introduction

More often than not, people ignore phone calls that have a random number on it when the call starts talking about politics. This study attempts to see if we can see if getting a call is a good replacement to listening to the call entirely because of bias with the treatment of listening to the call.

This study was conducted to see how effective these phone calls are and if they really do make an impact on society. People who took part in the experiment were called at random and were asked to vote in the coming election. Researchers recorded the results of this question. This paper consists of two types of data: experimental and non-experimental. Experimental data is data where researchers can randomly assign people to get the phone call, and non-experimental data is where researchers do not have control of who is assigned to the treatment. This is seen in the second data set where people selected themselves into the treatment. Also, balance tables ensure that the means of the two groups, treatment and control, are the same so that the control group is a good counterfactual for the treatment.

Using regressions, we can see the effect that each treatment group has on the outcome of voting. For each treatment groups, I ran two regressions. The first regression of each group looks at the treatment effect of each treatment groups on the outcome of voting. The second regression was the same one as the first, but with covariates. Adding covariates to the regression is important because they are correlated to both the treatment effect and the outcome. Also, covariates significantly influence the treatment effect. The other method used was Instrumental Variables, which sees if getting a call is a good substitute for listening to the call.

When running regressions with covariates, we see that as we add more covariates the experimental treatment group stays the same, but the non-experimental treatment group changes drastically. This is because there is something different in the non-experimental treatment group. Individuals selected themselves into the treatment group by choosing to listening to the entire phone call, which is selection bias. Moreover, we want to know why they listened to the phone call. An omitted variable makes the treatment of the experimental and non-experimental groups different. Due to this bias, implementing instrumental variables will be a good way to clear selection and omitted variable bias. However, we see that implementing instrumental variables only gives the local average treated effect for the group of compliers.

Data

The data was gathered in 2002, during the time of a midterm election in Iowa and Michigan and consists of 101,062 observations. For the sake of this experiment, the person who picked up the phone counts as the voter we are looking at. There may be another person in the household with the ability to vote, but this experiment will only focus on the person who picked up the phone. Households were randomly assigned to get a phone that asked them to vote in the coming midterm election. People who didn't answer the call were in the control group since they were not affected by the call; they are a representation of how people act without getting a call to vote. The first treatment group is the group that received a call. A part of the experimental data is to look at if simply getting this call would make people more likely to vote in the coming election. Many people assigned to get a call didn't listen to the call, nor did they have to be in the assignment. The non-experimental data consists of people selected themselves into the treatment group. The treatment in the non-experimental group is comprised of people who not only got the call but also stayed and listened to the entire thing. (Arceneaux et al, 2006)

Analyzing the effect of people on getting a call would give a significant amount of bias. Simply getting a call is not telling if people are more likely to vote or not. In the study, there are other variables analyzed to see if these variables had an overwhelming affect on the results. The main variables that looked at are the voter's gender (for simplicity only male and female), past voting history, if the voter is a newly registered voter, and his/her age. We believe that a person's willingness to vote is based on these factors also. Gender influences a person's likeliness to vote as one gender votes more than the other. In this study, a binary variable called female is used to state gender; 0 means the voter is male and 1 means the voter is female. Past voting history is crucial to look into because people who voted consistently in the past are more likely to vote in the 2002 election. Specifically, if the person voted in 1998 and in 2000, which are indicated by separate variables. Both of these are also binary since a person voted in the previous elections or didn't vote. The next variable is if the person is a newly registered voter or not, as it will affect voting in the nearest election. Age is an important variable because people who are older are more likely to vote than people who are younger. Looking at these factors, the data shows if people are more likely to vote in the 2002 election.

Before analysis of the data can happen, the variables must be consistent across the treatment and control. Creating a balance table will allow us to check if they are consistent or if they show large disparities. It also shows the p-value, which is an indicator of statistical significance.

Table 1: Balance Table for Experimental Treatment

	Control	Treatment	Difference	
	mean	mean	b	p
Age of Voter	55.90	55.51	0.385*	0.0205
If Voter is Female	56.2%	55.6%	0.007	0.1205
Newly Registered Voter	4.7%	5.1%	-0.004*	0.0498
If Voter voted 2000	73.4%	72.9%	0.005	0.1644
If Voter voted 1998	57.3%	57.1%	0.001	0.7711
Observations	85986	15076	101062	

One asterisk on the estimate means that the variable is significant on the 90% confidence interval. First for the experimental treatment, the values of the mean are relatively similar across the board. The average age of people receiving the call does not have a large disparity and the averages of the other variables does not have large disparities either. The control is a good counter factual to the treatment group. Moreover, the table shows a small p values on two variables, age of voter and newly registered voter, which means that these two variables are significant.

Table 2: Balance Table for Non-experimental Data

	Control	Treatment	Difference	
	mean	mean	b	p
Age of Voter	55.669	58.261	-2.59***	0.00
If Voter is Female	56.1%	56.7%	-0.01	0.31
Newly Registered Voter	4.8%	4.4%	0.00	0.09
If Voter voted in 2000	73.0%	77.5%	-0.05***	0.00
If Voter voted in 1998	57.0%	61.2%	-0.04***	0.00
Observations	94218	6844	101062	

In the non-experimental data, many of the variables have large disparities. The age of voters is higher on average in the treatment than in the control. The three stars indicate that the estimate is significant at the 99% interval. Also, the variables on previous voting history also show drastic differences. The people in the treatment are more likely to have voted in a previous election on average than people in the control group. The p values of these variables are very small, meaning that these variables are very significant. There isn’t a large disparity with the variables indicating gender and if the person is a newly registered voter. Also, these two variables have large p-values, showing that they aren’t significant. Since the means of many of the variables are not close to each other, the control does not offer a good counter factual to the treatment.

Methods

Regressions are used to see the treatment effect with each of the treatment groups. In these regressions, we will also add covariates to see if the there are other factors that make people more likely to vote. However, since we believe that there is bias within the non-experimental treatment group, we use the experimental treatment group as a substitute to see if we can see the effect of the get out the vote on the population.

The first thing is to run a regression seeing the effect of the treatment by itself with the experimental data. The first regression looks like this:

$$Equation\ 1: Votein2002_i = B_0 + B_1 AssignedCall_{\mathrm{i}}$$

The equation 1 regression would tell us the treatment effect of getting a call on the vote in 2002, which is what B_1 tells us. B_0 tells us the control, how likely people are to vote without getting the phone call. The variable Votein2002 for individual i is the outcome of voting, and AssignedCall is the assignment to getting a call. However, there may be more factors to voting other than simply getting a phone call. We get the following equation once we add more covariates that are likely to be correlated to the outcome of voting in 2002:

$$Equation\ 2:$$

$$Votein2002_i = B_0 + B_1 AssignedCall_i + B_2 Age_i + B_3 Female_i + B_4 NewVoter_i + B_5 Votein2000_i + B_6 Votein1998_i$$

After adding the important covariates in, this equation 2 should provide a good estimate of getting the phone call on the population. The new covariate of Age shows the age of the voter. Female is a binary variable that shows if the voter is female or male. NewVoter is a binary variable on if the person is a newly registered voter or not. Votein2000 and Votein1998 are both binary variables on previous voting history. The equation 2 regression shows if the treatment effect of getting a call, B_1, is changed when adding covariates.

We wish to see the treatment effect of listening to the call on the outcome of voting in 2002. Running the regression for the non-experimental treatment group looks like this:

$$Equation\ 3: Votein2002_i = B_0 + B_1 Treatment_i$$

The variable Treatment is the treatment effect in question: if listening to the call effects an individual's outcome in voting in the 2002 election. Equation 3 is the main regression we want to see; what the effect of these calls on the population as a whole. The same problem comes up as the first one when only running equation 3; there are other variables that might also be correlated with the outcome other than listening to the call. To combat this, we add more variables. We get a regression that looks something like this:

$$Equation\ 4:$$

$$Votein2002_i = B_0 + B_1 Treatment_i + B_2 Age_i + B_3 Female_i + B_4 NewVoter_i + B_5 Votein2000_i + B_6 Votein1998_6$$

Equation 4 demonstrates the effect of adding covariates on our B_1. Purposely adding the same variables in this regression is essential because we also want to know the effect of listening to the whole call versus the effect of simply getting the call.

Equation 3 is an important regression because it shows the effect of listening to the call on voting. However, the estimate we get from equation 3 is likely to be biased because people chose themselves into the treatment group. Another method to attempt to see the treatment effect of listening to the call on an unbiased subpopulation is to use an instrumental variable to take its place. This instrument is getting a call because we think getting a call and listening to the call are correlated. First, it is essential to get the first stage effect to see the number of compliers within the treatment group.

$$Equation\ 5: Treatment_i = \emptyset_0 + \emptyset_1 GettingCall_i + E_i$$

The $\emptyset_1$is the compliance rate of people in the who got the call and went through of the treatment of listening to the call. The next step is to use the reduced effect, the treatment effect in equation 1. The treatment effect of people getting the call. However, only a small number of people complied with the assignment to treatment. Dividing the reduced effect, the treatment of getting a call, by the first stage, compliance rate, will help us get the treatment effect for the compliers.

Results

In both experimental and non-experimental treatment groups, running regressions shows the respective treatment effect on the outcome of voting in 2002. In a general overview, the treatment effect of both are positively correlated to the outcome, but listening to the call all the way through makes people more likely to vote in the coming election than simply getting the call. This is due to the fact that there is selection bias in the treatment of the non-experimental data. Also, there is omitted variable bias because there is something inherently different about the people who selected themselves into the treatment.

Table 3 reports estimated coefficient from equation 1 and 2. This table has the independent variables on the y-axis and the effect of each of the variables on dependent variable.

In this table, more variables are added to see what the effect of these variables is on the treatment effect of getting the call. Also, three asterisks on the estimates denotes that the estimate is significant at the 99% confidence interval.

Table 3: Impact of Getting the Call on Outcome of Voting

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Vote in 2002	Vote in 2002	Vote in 2002	Vote in 2002	Vote in 2002	Vote in 2002
Getting the Call	0.012***	0.014***	0.014***	0.015***	0.016***	0.014***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Age of Voter		0.006***	0.006***	0.006***	0.003***	0.001***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
If Voter is Female			-0.032***	-0.032***	-0.027***	-0.024***
			(0.003)	(0.003)	(0.003)	(0.003)
Newly Registered Voter				-0.224***	0.130***	0.155***
				(0.007)	(0.007)	(0.007)
If Voter voted in 2000					0.533***	0.400***
					(0.003)	(0.004)
If Voter voted in 1998						0.276***
						(0.003)
Constant	0.594***	0.283***	0.290***	0.328***	0.029***	0.073***
	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	101,062	101,062	98,454	98,454	98,454	98,454
R-squared	0.000	0.046	0.054	0.063	0.250	0.304

Initially, the estimated effect of getting a call is positive (B_1), that people who get the call are more likely to vote in the 2002 election. But, the treatment effect, B_1, stays the same. Adding more variables doesn't have an effect on the treatment effect. This is because of the fact that the treatment was assigned randomly, meaning there is low bias in the estimate. Also, there is a large number of observations in the data set. Moreover, the variables are significant due to a low standard error, this further bolsters the fact that this data set is representative of the population.

Table 4 estimates the coefficients from running the regressions of equation 3 and 4. Table 4 has independent variables on the y-axis of the graph and effect of each of the variables on the dependent variable of voting in 2002. The treatment effect in this graph is listening to the call all the way through. The three asterisks on the estimates denotes the fact that the estimates are significant at the 99% confidence interval.

Table 4: Impact of Listening to the Call on Voting in 2002

	(1)	(2)	(3)	(4)	(5)	(6)
				Voting	Voting	Voting
VARIABLES	Voting 2002	Voting 2002	Voting 2002	2002	2002	2002
Listening to the Call	0.082***	0.068***	0.059***	0.060***	0.048***	0.046***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
Age of Voter		0.006***	0.006***	0.006***	0.003***	0.001***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
If Voter is Female			-0.032***	-0.032***	-0.027***	-0.024***
			(0.003)	(0.003)	(0.003)	(0.003)
Newly Registered Voter				-0.224***	0.130***	0.155***
				(0.007)	(0.007)	(0.007)
If Voter voted in 2000					0.533***	0.399***
					(0.003)	(0.004)
If Voter voted in 1998						0.276***
						(0.003)
Constant	0.590***	0.282***	0.290***	0.328***	0.029***	0.074***
	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	101,062	101,062	98,454	98,454	98,454	98,454
R-squared	0.002	0.047	0.055	0.064	0.251	0.304

In table 4, adding more variables decreases the treatment effect of listening to the phone call. What this means is that the covariates have influence with the treatment variable, demonstrating that the covariates are correlated with the with the treatment effect and the outcome.

In the non-experimental treatment group, people were allowed to select themselves into the treatment group. The fact that people are allowed to select into different groups, control or treatment, demonstrates the selection bias in the non-experimental data. Selection bias is the ability for people in either the treatment or the control group to choose into another group. This is why we see a drastic decrease when we add more variables into table 4. Moreover, people who listen to political phone calls until the end are probably different than those who simply were assigned to get a phone call. This difference demonstrates an omitted variable in our regression, which bias our estimates in the regression. The reason why people listened to the call is the omitted variable, but this variable isn’t something we can observe. For example, we can’t measure political fervor on a quantitative scale.

Due to the selection and omitted variable bias, including the instrumental variable will demonstrate the treatment effect on the subpopulation better than the regression in table 4. To first see the effect of this instrumental variable, we need to get the reduced effect, which is the treatment effect of getting the call on voting in 2002 in the population. This is the treatment effect in equation 1: the B_1. The first stage is the treatment effect of equation 5, and we get the first stage estimate, which is $\emptyset_1$. This estimate is the compliance rate of people in the treatment group. We want to scale the reduced form to be representative of the group of compliers. To do this, we divide the reduced form by the first stage to get the treatment effect on the group of compliers. This get the instrumental variable effect; the treatment effect of getting the call for people who complied with the experiment. Table 5 shows the treatment effect for the compliers in the experiment. Because there is something different between the compliers and the non-compliers, this table is only representative of the subpopulation of compliers and not the entire population.

Table 5: Effect of the Instrumental Variable :Getting a Call

	(1)
	Vote in
VARIABLES	2002
Listening to the call	0.026***
	(0.010)
Constant	0.594***
	(0.002)
Observations	101,062
R-squared	0.001

Conclusion

This paper looks at a handful of statistical methods to see if we can see an unbiased treatment effect of the non-experimental data. Using randomization, we can eliminate bias within the regression, but wonder if the variable used to eliminate bias has an effect on people. The treatment effect we care about is seeing if people listening to the phone call will make people more likely to vote. We wish to see if it is plausible to include an instrumental variable of getting a call in place of the treatment in the call. This would allow the best of both worlds, since the instrumental variable is randomized and it allows to see the effect of listening to the call. What we find is that the instrumental variable estimate is not represented of the entire population.

The main methods used in this study are regressions. Regressions demonstrate the treatment effect of each treatment group and show the effect of adding covariates into the regressions. However, both the regression gives a biased estimate of the population and subpopulation. To fix this, we use the instrumental variable estimate, which shows the treatment effect on the subpopulation of people who listened to the call.

The experimental data consisted of a randomized sample, which means that there is no bias in the estimates. Meaning that the experimental data is a good sample of the population, but doesn't offer us the effect of the actual treatment of listening to the call. Furthermore, when running the regression containing the covariates, the covariates have a minimal change on the coefficient in the experimental data. When running the same regressions with the non-experimental data, we find another story. The regressions of the non-experimental data show that the variables are highly correlated with both the treatment effect and the outcome. Adding more covariates into the regression equation makes the coefficent on the treatment variable decreases significantly. The major difference in the experimental data was that people had the ability to choose themselves into the treatment group. When people are allowed to choose themselves into the control or treatment group is known as selection bias. Also, there is inherently different about people who select themselves into the treatment, because of this there is omitted variable bias. This variable is unobservable because it is hard to characterize what makes people listen to a political phone. We can add as many covariates as we want and won't be able to get rid of this bias. However, both of these regressions contain bias, implementing instrumental variables helps to eliminate the bias. The treatment effect of the instrumental variable only shows the treatment effect of a subpopulation of people who listened to the phone call and is not indicative of the entire population.

The study is important because it shows the benefits and downsides of multiple experimental methods. The benefits of having a randomized treatment group is there is no bias, but also the downside that it doesn't have the correct treatment group. Running a regression with the effect of listening to the call on the outcome of voting has the downside of having selection and omitted variable bias. Using an instrumental variable has the benefit of taking care of this bias, but the downside of showing the effect of the treatment on a subpopulation. All this demonstrates that statistical methods can never be 100% accurate and contain upsides and downsides.

References

Arceneaux, Kevin, et al. “Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment.” Political Analysis, vol. 14, no. 1, 2006, pp. 37–62. JSTOR, www.jstor.org/stable/25791834.