CS 112 - Final Project

Statistical Replication and Extension

-

Ajaydip Singh and Petter Hallqvist

17th of December 2019

Research Paper:
Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions (Alvarez, Levin, Pomares and Leiras, 2007)

Link to paper:

https://www-cambridge-org.ccl.idm.oclc.org/core/services/aop-cambridge-core/content/view/A96562BA348433894BED1F2BA8BC3761/S2049847013000022a.pdf/voting_made_safe_and_easy_the_impact_of_evoting_on_citizen_perceptions.pdf

Link to original data and code:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/24896

From: Ajaydip Singh and Petter Hallqvist

To: R. Michael Alvarez, Ines Levin, Julia Pomares and Marcelo Leiras

Regarding: Comparison of voter experience using traditional voting and e-voting techniques, replication and extension of Alvarez et. al (2013) results using genetic matching with propensity scores.

Summary

This memo is a replication of propensity score matching analysis conducted by Alvarez et. al (2013), and extends on their methods using genetic matching. Alvarez et. al (2013) originally collected observational data on voter experience of using e-voting machines in comparison to traditional voting. ***Hence, the decision question explored in relation to this paper is if e-voting should be widely implemented in Argentina in order to improve voter experience.*** *Our replication confirms the conclusions of the authors, that e-voting does improve voter experience. However, using genetic matching methods we find evidence for a higher causal effect on voter experience, higher voter speed and more concern about ballot secrecy than the original paper suggests.* ***Our policy advice is to implement e-voting in Argentina but we emphasize in designing the e-voting system in a way that increases ballot confidence among citizens****. A related question explored in this memo is whether using the MatchIt or Matching libraries is more appropriate for genetic matching methods. The conclusion of this memo is that the Matching library is more appropriate - due to its ease of use, model transparency, availability of a genetic matching function and MatchBalance function. The main replication done is of Table 1 in this paper measuring pre and post matching balance, however, we also included a replication of Table 4 in this paper to show the significance of our analysis.*

Structure of Memo

This memo aims at replicating the results of Alvarez et. al (2013), and extending their analysis to obtain more accurate estimations of the treatment effect. By doing this, we are informing the broader decision question: Should e-voting be widely implemented in Argentina in order to improve voter experience?

The structure of this paper includes the following: Section 1- A general introduction to Alvarez et. al (2013) (including data collection and statistical analyses methods); Section 2 - Rationale of replication and extension method chosen; Section 3 - Replication of Table 2 of their paper using propensity score matching; Section 4 - Extension of results through genetic matching; Section 5 - Recommendations in relation to the Decision Question; Section 6 - Conclusion and discussion.

1) Introduction to the paper

Alvarez, Levin, Pomares, & Leiras (2013) study focused on voter perception and evaluations of e-voting (EV) relative to traditional voting (TV) in Salta, Argentina. The data collection method involved using questionnaires in 36 different polling stations (N = 1,502, EV = 887, TV = 615). Each polling station either administered an EV or TV system, but this assignment was not randomized. This implies that the EV and TV groups are confounded and vulnerable to model dependence if regression adjustment methods were used. Alvarez et. al (2013), hence, used propensity score matching (PSM) – using *MatchIt* library in R – so that the results are less sensitive to model dependence, and a more reliable causal estimate could be produced.

Specifically, Alvarez et. al (2013) estimated the Average Treatment effect of the Treated (ATT), used a caliper of 0.05 (a condition for the match to be lower than 5% of the standard deviation of the estimated propensity score) and found a unique match for each unique EV unit based on the matched covariates i.e. did matching without replacement. The matching process led to 311 observations being discarded. *Table 1* shows the match balance Alvarez et. al (2013) obtained originally using PSM.

2) Rationales of Methods used

PSM aims to find a randomized experiment within observational data. This is done by reducing covariate imbalance using a one-dimensional proxy for covariate distribution. Balancing covariates across treatment and control groups satisfies the assumption of ignorable treatment assignment, meaning that it is important to balance on all variables known "to be related to both the treatment assignment and outcome" (Stuart 2010).

However, critics have described PSM as suboptimal because even though it simplifies the dimensionality of the matching problem, the propensity scores does not capture the complexity of the covariate distribution. It is also argued that PSM are only useful in theory when PS are known and not in practice when they are estimated. Hence, in our extension we used covariate matching in addition to PSM, and a genetic algorithm (the GenMatch function in the *Matching* library) to select optimal matches and drop fewer observations in the process.

	Before matching (N = 1,475)				After matching (N = 1,164)				
	EV	TV	Diff.	p-value*	EV	TV	Diff.	p-value*	% Imp.
Age group (1–5)	2.5	2.4	0.0	0.55	2.5	2.5	0.0	1.00	100%
Education (1–8)	4.8	4.1	0.6	0.00	4.2	4.2	0.0	0.72	98%
White collar (%)	30.3	27.6	2.7	0.29	29.2	28.4	0.9	0.80	68%
Not full time worker (%)	27.7	33.5	−5.8	0.02	30.8	32.0	−1.2	0.80	79%
Male (%)	49.7	49.1	0.6	0.87	49.0	49.0	0.0	1.00	100%
Technology count (1–6)	4.2	3.9	0.3	0.00	4.0	3.9	0.1	0.59	76%
Political information (1–4)	1.5	1.3	0.2	0.00	1.4	1.3	0.0	0.55	77%

*Table 1 - From Alvarez et. al (2013) (Table 2) showing balance of individual covariates in treatment and control groups before and after matching. Units were matched on propensity scores which were estimated using the covariates: age.group, age.group^2, age.group^3, age.group*educ, age.group*tech, educ, educ^2, tech, tech^2, pol.info, educ*pol.info, age.group*pol.info, tech*pol.info, white.collar, not.full.time, and male. *p-values of Kolmogorov-Smirnov tests (ordinal variables), and difference in proportions tests (binary variables)*

3) Replication of The Matching Procedure in Alvarez et. al (2013)

Using the code provided by Alvarez et. al (2013) we replicated the matching procedure using Matchit (see table 2). We were unable to replicate the % Imp. variable because the authors did not provide the code and did not explain how the variable was calculated. The difference in mean EV and TV for the various covariates we obtained after matching were slightly different from the original results. However, the p-values for the Kolmogorov-Smirnov and difference in proportions test that we obtained after matching were consistently lower. The lower p-values indicate that we achieved a poorer balance when replicating their results.

	Before matching (N = 1,475)				After matching (N = 1,142)			
	EV	TV	Diff.	p-value	EV	TV	Diff.	p-value
Age group (1-5)	2.5	2.4	0.0	0.57	2.5	2.5	0.0	1.00
Education (1-8)	4.8	4.1	0.6	0.00	4.2	4.2	0.0	0.49
White collar (%)	30.3	27.6	2.7	0.29	29.4	27.3	2.1	0.47
Not full-time worker (%)	27.7	33.5	-5.8	0.02	30.6	32.1	-1.5	0.61
Male (%)	49.7	49.1	0.6	0.87	49.0	49.8	-0.9	0.81
Technology count (1-6)	4.2	3.9	0.3	0.00	4.0	3.9	0.1	0.31
Political information (1-4)	1.5	1.3	0.2	0.00	1.4	1.3	0.0	0.63

Table 2 - Replication of the matching procedure in Alvarez et. al (2013) using the same MatchIt package in R.

We also replicated the matching procedure using the *Matching* library (see *Appendix 2*) (Sekhon 2019) . Our reasoning behind this is related to the subsequent extension of Alvarez et. al (2013)'s matching using genetic matching. While the *MatchIt* library enables genetic matching, it does so using the *GenMatch()* function of the *Matching* library set at unknown default

parameters, and is hard to interact with (Ho et. al 2011). Hence, we prefer using the *Matching* library when performing genetic matching to manually be able to change arguments. Furthermore, we want to include the replication of PSM results using the *Matching* library so as to effectively compare it with our extension using genetic matching.

4) Extension of results through Genetic Matching

Genetic Matching applies a genetic algorithm found in the *rgenoud* library in R to determine optimal matches for balance on observed covariates between treatment and control groups. The default parameter which the genetic algorithm optimizes for is maximizing p-values from paired t-tests and Kolmogorov-Smirnov tests, but this can be changed through the fit.func parameter of the GenMatch function in the *Matching* library (Sekhon 2019).

When we performed Genetic Matching we used all of the covariates used in the original paper, and added I(tech^3) and I(white.collar^2) and propensity scores (*see Appendix 3 for code*). The decision of adding these additional covariates was made based on the recommendations made by Stuart (2010) on the importance of balancing on variables influencing the outcome. We consider it a high likelihood that technical ability and comfort using technology professionally would influence voter experience using e-voting, which is why we wanted to ensure these covariates were well balanced.[1]

Originally we included a 0.05 caliper in the *GenMatch* function (like the original paper), however we found that this led to significant numbers of units dropped (58% of control units and 48% of treated units). We hypothesize that this may be because the function tried to match units very closely to each other – almost exact matching – which is why so many units were dropped. When we removed the 0.05 caliper, we only dropped 208 control units, which is an improvement from the original study while p-values generally remained very high and the lowest p-value was 0.3, suggesting the balance that we achieved was better than what Alvarez et. al (2013) achieved (*See Table 3)*. We did all our analysis blind to outcomes as Rubin recommends i.e. we attempted different genetic matching models and used only p-values provided from *MatchBalance* and the number of observations dropped to assess the balance we achieved before we elected one method that we found optimal (*see other genetic matching attempts in appendix 4*).[2]

---

[1] #variables: Based on the recommendations of *Stuart (2010)* we chose to include the covariates tech^3 and white.collar^2 to strengthen the assumption of ignorable treatment assignment by balancing on variables influencing outcomes - citizen experience with e-voting.

[2] #algorithms: We choose a genetic matching model that achieves a high balance and leads to fewer dropped observations. We experimented with different parameters such as the caliper, replace and ties. We also attempted genetic matching with and without propensity scores. After trying multiple methods, we chose genetic matching with propensity scores with two additional covariates of our choice.

	Before matching (N = 1,475)				After matching (N = 1,732)			
	EV	TV	Diff.	p-value	EV	TV	Diff.	p-value
Age group (1-5)	2.5	2.4	0.0	0.57	2.5	2.5	0.0	1.00
Education (1-8)	4.8	4.1	0.6	0.00	4.8	4.8	0.0	0.92
White collar (%)	30.3	27.6	2.7	0.29	30.3	30.3	0.0	1.00
Not full-time worker (%)	27.7	33.5	-5.8	0.02	27.7	26.7	1.0	0.67
Male (%)	49.7	49.1	0.6	0.87	49.7	49.5	0.1	1.00
Technology count (1-6)	4.2	3.9	0.3	0.00	4.2	4.2	0.0	0.90
Political information (1-4)	1.5	1.3	0.2	0.00	1.5	1.5	0.0	0.98

Table 3 - Extension of the matching procedure in Alvarez et. al (2013) using genetic matching with propensity scores.

5) Recommendations in relation to Causal Question

The causal estimates of voter perception on e-voting that Alvarez et. al (2013) obtained before and after the matching procedure are summarized in *Table 4* below.

	Before matching (N = 1,475)					After matching (N = 1,164)				
	N	E-Voting (%)	Traditional Voting (%)	Diff.	p-value*	N	E-Voting (%)	Traditional Voting (%)	Diff.	p-value*
Select candidates electronically	1,388	83.8	53.4	30.4	0.000	1,101	82.7	54.1	28.6	0.000
Evaluation of voting experience	1,460	46.3	21.3	25.0	0.000	1,151	45.6	20.9	24.7	0.000
Ease of voting procedure	1,469	33.6	11.5	22.1	0.000	1,159	32.5	11.9	20.6	0.000
Agree substitute TV by EV	1,409	84.1	62.4	21.7	0.000	1,114	82.4	63.3	19.1	0.000
Elections in Salta are clean	1,284	58.0	41.0	17.0	0.000	1,022	57.6	41.5	16.0	0.000
Sure vote was counted	1,418	86.3	77.0	9.3	0.000	1,117	85.7	77.0	8.8	0.000
Qualification of poll workers	1,416	85.1	76.2	8.9	0.000	1,123	84.5	76.0	8.5	0.000
Speed of voting process	1,443	84.1	80.8	3.2	0.130	1,137	83.2	80.7	2.5	0.306
Confident ballot secret	1,431	77.1	84.5	−7.4	0.001	1,133	76.9	84.3	−7.4	0.002

*Table 4 - Causal estimates of voter perception on EV obtained by Alvarez et. al (2013) before and after matching. *Test of difference in proportions.*

As can be seen in *Table 4* above, Alvarez et. al (2013) concluded that in general the electronic voting system experience in Salta is evaluated more favorably than voters who used the traditional voting system. Alvarez et. al (2013) also commented that they were surprised to see only a small difference in the speed of the voting process, and further expressed that there are more concerns about the ballot secrecy in e-voting in contrast to traditional voting. However, from our extension of their matching procedure, we obtained the following causal estimates with our more balanced post-matching population (*see Table 3)*:

	Before matching (N = 1,475)					After matching (N = 1,732)				
	N	E-Voting (%)	Traditional Voting (%)	Diff.	p-value	N	E - Voting (%)	Traditional Voting (%)	Diff.	p-value
Select candidates electronically	1388	83.8	53.4	30.4	0.000	1631	83.8	58.0	25.9	0.000
Evaluation of voting experience	1460	46.3	21.3	25.0	0.000	1708	46.3	16.1	30.2	0.000
Ease of voting procedure	1469	33.6	11.5	22.1	0.000	1726	33.6	10.2	23.5	0.000
Agree substitute TV by EV	1409	84.1	62.4	21.7	0.000	1651	84.1	65.2	18.9	0.000
Elections in Salta are clean	1284	58.0	41.0	17.0	0.000	1517	58.0	40.9	17.0	0.000
Sure vote was counted	1418	86.3	77.0	9.3	0.000	1667	86.3	79.1	7.3	0.000
Qualification of poll workers	1416	85.1	76.2	8.9	0.000	1671	85.1	76.4	8.7	0.000
Speed of voting process	1443	84.1	80.8	3.2	0.130	1704	84.1	78.1	5.9	0.002
Confident ballot secret	1431	77.1	84.5	-7.4	0.001	1679	77.1	86.1	-9.0	0.000

Table 5 - Causal estimates of voter perception on EV obtained before and after matching using genetic matching with propensity scores .

Our causal estimates show a higher positive voter experience for e-voting than what Alvarez et. al (2013) originally estimated (30.2 in contrast to 24.7). We also found a larger difference in the speed of voting between the two voting systems (we found a difference of 5.9 between EV and TV while Alvarez et. al (2013) found 2.5). In relation to ballot secrecy, our estimates suggest that voters were less confident about ballot secrecy for e-voting in relation to traditional voting systems. Even though we did not do the post regression adjustment and sensitivity analysis that Alvarez et. al (2013) did, our extensions of their matching procedure provide strong evidence that overall e-voting was more favorable for people who used the e-voting system in contrast to people who used the traditional voting. Our policy advice is to implement e-voting systems as they are generally more preferable and less time consuming. However, we emphasize that the e-voting systems should be designed to increase confidence in ballot secrecy so that less people are skeptical about it.

6) Conclusion and discussion

In conclusion, while propensity score matching improves observational study designs and reduces the effect of model dependence, it is only suboptimal. In this memo we recommend using genetic covariate matching rather than reducing the dimensionality of the problem to a single dimension as is done in PSM. We also argued that the balance we achieved was better than Alvarez et. al (2013) as it led to lower dropped observations and higher p-values. A higher balance would produce more reliable and accurate causal estimates, however, it is important to keep in mind that a higher balance often leads to more observations being dropped and therefore one must consider the tradeoff between the two when choosing an optimal method – it also depends on the size of the dataset.

Our replicated results confirmed the conclusions drawn by Alvarez et. al (2013). However, we obtained a larger causal estimate for the effect of e-voting on voter experience, and

we measured a more significant difference in voting speed than the original paper. This can largely be attributed to the higher balance we were able to achieve by using genetic matching. The scope of this replication was limited, and did not cover the sensitivity analysis which Alvarez et. al (2013) carried out.[3]

A problem we encountered using the *MatchIt* library was the inability to consistently assess the balance achieved after the matching procedure. In the *Matching* library, this problem did not arise because the *MatchBalance* function conveniently provides both t-test and Kolmogrov-Smirnov p-values after the matching procedure is done (Sekhon 2019). Furthermore, we attempted to perform genetic matching through *MatchIt* library but could neither change the default parameters beyond *method = "genetic"* nor discover what the default parameters for *population, wait.generations* and *max.generations* were. The R Documentation file for *MatchIt* suggests that we should be able to change these parameters, but even after extensive research we did not manage to do so (Ho et. al 2011) (see *appendix 5)*. The lack of ability to manually manipulate variables and understanding models are obstacles for researchers trying to both optimizing balance for their matching models and interpreting them. This is a pertinent problem for conducting inference analysis and for academic peer-review of the validity of that analysis. Based on these factors, going forward we recommend using the *Matching* library instead of the *MatchIt* library for statistical matching and causal inference analysis.

WORD COUNT: 1931

[3] #studyreplication - This memo replicated data processing by *Alvarez et. al (2013),* commented on the method of data analysis used in the *MatchIt* library, and extended the analysis using genetic matching with the *Matching* library. With higher p-values for balance and less dropped observations, we improved the treatment effect estimate of *Alvarez et. al (2013)*

Bibliography

Alvarez, R. Michael. Levin, Ines. Pomares, Julia. Leiras, Marcelo. (2013). "Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions" *The European Political Science Association.* Vol 1, No. 1, 117–137 June 2013

Ho, Daniel E. Imai, Kosuke. King, Gary. Stuart, Elizabeth A. (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Cran R-Project.* June 28th 2011.

Sekhon, Jasjeet Singh. (2019). "Package 'Matching' " *Cran R-Project.* April 7th 2019.

Stuart, Elizabeth A. (2010) "Matching methods for causal inference: A review and a look forward" *Statistical Science.* 2010 February 1; 25(1): 1–21.

Appendix

Appendix 1 - Replication of Alvarez et. al (2013). MatchIt Library - PSM

Appendix 2 - Matching library Replication of Results

Code: https://gist.github.com/Ajaydip-Singh/ea68e086c1beaf6cf43c4358a6374b8d

	Before matching (N = 1,475)				After matching (N = 1,142)			
	EV	TV	Diff.	p-value	EV	TV	Diff.	p-value
Age group (1-5)	2.5	2.4	0.0	0.57	2.5	2.5	0.0	1.00
Education (1-8)	4.8	4.1	0.6	0.00	4.2	4.2	0.0	0.49
White collar (%)	30.3	27.6	2.7	0.29	29.4	27.3	2.1	0.47
Not full-time worker (%)	27.7	33.5	-5.8	0.02	30.6	32.1	-1.5	0.61
Male (%)	49.7	49.1	0.6	0.87	49.0	49.8	-0.9	0.81
Technology count (1-6)	4.2	3.9	0.3	0.00	4.0	3.9	0.1	0.31
Political information (1-4)	1.5	1.3	0.2	0.00	1.4	1.3	0.0	0.63

Appendix 3 - Extension of Alvarez et. al (2013). Matching Library - Genetic Matching

Appendix 4 - Other genetic matching attempts

1. Covariates matched on: *age.group, age.group^2, age.group^3, age.group*educ, age.group*tech, educ, educ^2, tech, tech^2, pol.info, educ*pol.info, age.group*pol.info, tech*pol.info, white.collar, not.full.time, and male*

Replication with GenMatch – min p value: 1 ---> ties = F, replace = F, estimand = ATT, M = 1, caliper = 0.05

	Before matching (N = 1,475)				After matching (N = 574)			
	EV	TV	Diff.	p-value	EV	TV	Diff.	p-value
Age group (1-5)	2.5	2.4	0.0	0.57	2.3	2.3	0.0	1.00
Education (1-8)	4.8	4.1	0.6	0.00	4.3	4.3	0.0	1.00
White collar (%)	30.3	27.6	2.7	0.29	24.4	24.4	0.0	1.00
Not full-time worker (%)	27.7	33.5	-5.8	0.02	28.6	28.6	0.0	1.00
Male (%)	49.7	49.1	0.6	0.87	45.6	45.6	0.0	1.00
Technology count (1-6)	4.2	3.9	0.3	0.00	4.1	4.1	0.0	1.00
Political information (1-4)	1.5	1.3	0.2	0.00	1.2	1.2	0.0	1.00

2. Covariates matched on: *age.group, age.group^2, age.group^3, age.group*educ, age.group*tech, educ, educ^2, tech, tech^2, pol.info, educ*pol.info, age.group*pol.info, tech*pol.info, white.collar, not.full.time, and male*

Replication with GenMatch - min p value = 0.47956 ---> replace=T, ties = F, estimand = ATT

	Before matching (N = 1,475)				After matching (N = 1,732)			
	EV	TV	Diff.	p-value	EV	TV	Diff.	p-value
Age group (1-5)	2.5	2.4	0.0	0.57	2.5	2.5	0.0	0.99
Education (1-8)	4.8	4.1	0.6	0.00	4.8	4.8	0.0	0.95
White collar (%)	30.3	27.6	2.7	0.29	30.3	30.6	-0.3	0.92
Not full-time worker (%)	27.7	33.5	-5.8	0.02	27.7	27.0	0.7	0.79
Male (%)	49.7	49.1	0.6	0.87	49.7	50.2	-0.6	0.85
Technology count (1-6)	4.2	3.9	0.3	0.00	4.2	4.2	0.0	0.95
Political information (1-4)	1.5	1.3	0.2	0.00	1.5	1.5	0.0	0.98

Dropped 447 controls, 0 treatments.

Appendix 5 - Matchit Library Genetic Matching Results

Code: https://gist.github.com/Ajaydip-Singh/629ae45dcdf5ed1c44b1a391e8b0ed4b

Team Contributions

Both: Since December 2nd we have been meeting every or every second day, reading replicating and extending Alvarez et. al (2013) for at least 30 minutes. Throughout this process, we have been reading up on statistical methods covered in class, and considering how we can best use these in our final project. On our own, we both played around with the GenMatch function to gain a better understanding of balancing factors with regard to our data. We both wrote the memo together, iteratively editing each other's text.

Ajaydip: Wrote the code enabling us to replicate and extend the analysis using the *Matching* library, as the code used by the authors was geared toward using outputs from the *MatchIt* library.

Petter: Did research on the strengths and limitations of the *MatchIt* library in comparison to the *Matching* library, and conducted a literature review informing the choice of covariates.[4]

[4] #differences - We set up a schedule where we were able to meet up over an extended period of time and consistently bounce ideas off each other, and approach the assignment iteratively and mindfully while keeping different schedules. Furthermore, to increase efficiency, we delegated work to complement each other's strengths of coding and comparing statistical analyses tools for academic purposes.