Minerva Schools @ KGI

Data Science And Analytics Department

Memo

To: Alexis Diamond, Head of the Data Science Department at MSKGI.

From: Fabian Okafor, Brian Wahome, Chief Data Scientists

Date: April 20th, 2018

Re: Analytical Extensions and Results Validation:The Unintended Consequences of Voter Persuasion Efforts by Daniel J. Hopkins.

EXECUTIVE SUMMARY.

Daniel J Hopkins; in his paper "The Unintended Consequences of Voter Persuasion Efforts" explored the effects of various campaign methods including canvassing, phone calls and mail on voting and voter turnout in the 2008 elections. Although this analysis was done mainly in a bid to review the importance of randomized experiments, the results of this analysis could be of importance to a large number of people; of which election campaigners are at the top of the list. Given this information, a potential audience for this paper is Barack Obama's campaign manager who wishes to know the effect of her campaign methods on voter turnout, voting and the outcome of the elections, and subsequently strategize future campaigns in favor of Obama. Framing this as a decision question;

"Given that you, as an election campaign manager for a general election in the US wishes to represent your client for a second term, and ensure your client's success in the upcoming election by either maximizing your client's votes or minimizing your opponent's votes, or

both, to whom should you direct your campaign strategies, and what campaign strategies should be best used?"

PAPER ABSTRACT

Hopkin's performed an analysis of data obtained from randomized field experiments in Wisconsin attempting to evaluate the effect of persuasive canvassing, phone calls and/or mail. While campaigns are designed to boost both turnout and support for the campaigning candidate, the results obtained were counter intuitive. The experiment relied on follow up surveys through one or more of the three listed methods. Randomization ensured the balancing of key and unobserved covariates in the experiment and given the large sample size(56000 observations with 12442 successful interviews conducted), the data can be assumed to be well balanced. The campaign techniques had a negative effect on the following three key variables:

- Follow up survey response rate.
- Likelihood to turn out and vote.
- Likelihood to support candidate(Backlash)

Statistical analyses were done to validate the balance of covariates among survey respondents for some of the campaign techniques. Canvassing for instance, was tested by conducting t-tests on both the control and treatment groups. The results were presented in Table 1 of the Persuasion paper. We replicated these results to validate the findings.

The evaluations above were investigating the candidate support aspect of campaigns and thus focused on the sample of 12442 who followed up on the surveys. If we however, were to zoom out and focus on the turnout, then the whole sample of 56000 qualifies since we have a

turnout score for each respondent. Evaluating the effect of treatment on the turnout score produced Table 3 in the paper. This was an OLS Estimate of the effect of treatment on the probability of turnout. At first glance, one might infer that Canvassing and mailing actually had positive effects on the turnout. But from our discussion above, we see the effect as negative. But why the variation in results? This is because of Selection Bias. We know that the treatment had different effects on different groups depending on their attitude towards the candidate. Those who supported them strongly for instance, were not likely to backlash. Those who were undecided, however, were less likely to turn out for the voting had they been pestered by the campaign treatments. Canvassing in general, turned off people unlikely to vote from answering the follow-up survey while it encouraged people who vote sometimes but not always. This means that in the survey sample, we have removed a disproportionate number of low-turnout voters who were canvassed and included a disproportionate number of moderate-turnout voters who were canvassed, thereby inducing a spurious association between canvassing and turnout. We will replicate the data in the next section to validate the findings by Hopkins.(H., & D. (2014, May 26).)

RESULTS REPLICATION AND VALIDATION.

Table 1: Balance Among Survey Respondents.(H., & D. (2014, May 26).)

	Mean			
	Canvass assigned	Canvass not assigned	P-value	N
Age	55.76	55.88	0.726	9,416
Black	0.017	0.018	0.671	12,442
Male	0.394	0.391	0.729	12,442
Hispanic	0.043	0.045	0.588	12,442
Voted 2002 general	0.242	0.232	0.163	12,442
Voted 2004 primary	0.390	0.371	0.031	12,442
Voted 2004 general	0.863	0.843	0.001	12,442
Voted 2006 primary	0.192	0.188	0.576	12,442
Voted 2006 general	0.634	0.600	0.000	12,442
Voted 2008 primary	0.429	0.406	0.011	12,442
Turnout score	3.263	3.149	0.005	12,442
Obama expected support score	47.36	47.95	0.100	12,440
Catholic	0.183	0.177	0.434	12,442
Protestant	0.467	0.455	0.181	12,442
District % Dem. 2004	54.66	54.86	0.353	12,440
District Dem. performance	58.01	58.18	0.374	12,440
District median income	46.26	45.94	0.155	12,439
District % single parent	8.19	8.28	0.212	12,439
District % poverty	6.22	6.40	0.127	12,439
District % college grads	19.79	19.58	0.279	12,439
District % homeowners	71.16	71.02	0.656	12,439
District % urban	96.64	96.96	0.099	12,439
District % white collar	36.31	36.29	0.882	12,439
District % unemployed	2.616	2.642	0.555	12,439
District % Hispanic	2.773	2.795	0.824	12,439
District % Asian	0.787	0.803	0.560	12,439
District % Black	1.849	1.878	0.759	12,439
District % 65 and older	22.82	22.80	0.921	12,439

We were not able to replicate the whole dataset because the dataset Hopkins made available was not full. We, however, were able to replicate the exact values for most of the variables though, including all P-values.

Replicated Results.

```
> round(rmatc,digits=3)
```

	Mean, C=1	Mean, C=0	P-value	Mean, C=1	Mean, C=2	P-value	N	N.survey
q_age	54.646	54.689	0.802	55.756	55.875	0.726	39187	9416
black	0.021	0.018	0.037	0.017	0.018	0.671	56000	12442
male	0.408	0.403	0.238	0.394	0.391	0.729	56000	12442
hispanic	0.054	0.056	0.355	0.043	0.045	0.588	56000	12442
vh_02g	0.206	0.204	0.523	0.242	0.232	0.163	56000	12442
vh_04p	0.329	0.329	0.943	0.390	0.371	0.031	56000	12442
vh_04g	0.830	0.831	0.910	0.863	0.843	0.001	56000	12442
vh_06p	0.154	0.160	0.052	0.192	0.188	0.576	56000	12442
vh_06g	0.551	0.550	0.786	0.634	0.600	0.000	56000	12442
vh_08p	0.356	0.351	0.254	0.429	0.406	0.011	56000	12442

.

.

.

```
> library(xtable)
```

Table 3: OLS estimates of treatment on probability of turnout.(H., & D. (2014, May 26)***.)***

	All subjects	Survey sample only
Canvass	0.003	0.015
	(0.004)	(0.008)
Phone call	-0.004	0.013
	(0.004)	(0.008)
Mail	0.001	-0.005
	(0.004)	(0.008)
Constant	0.664	0.726
	(0.004)	(0.008)
N	56,000	12,442

Standard errors in parentheses

Replicated results:

```
> texreg(list(all.sub,svy.sub),digits=3,stars=0.05)

\begin{table}
\begin{center}
\begin{tabular}{l c c }
\hline
 & Model 1 & Model 2 \\
\hline
(Intercept)      & $0.664^{*}$          & $0.726^{*}$ \\
                 & $(0.004)$    & $(0.008)$   \\
canvass          & $0.003$      & $0.015$     \\
                 & $(0.004)$    & $(0.008)$   \\
phonecall        & $-0.004$     & $0.013$     \\
                 & $(0.004)$    & $(0.008)$   \\
mail             & $0.001$      & $-0.005$    \\
                 & $(0.004)$    & $(0.008)$   \\
\hline
R$^2$            & 0.000        & 0.001       \\
Adj. R$^2$       & -0.000       & 0.000       \\
Num. obs.        & 56000        & 12442       \\
RMSE             & 0.472        & 0.440       \\
\hline
\multicolumn{3}{l}{\scriptsize{$^*p<0.05$}}
```

EXTENSIONS ANALYSIS.

As an extension of the analysis and results accumulated by Daniel J Hopkins in his paper, we wished to further make the decision problem more specific. By applying some polishes and shines on the previous decision problem, our new decision problem was narrowed down to the following;

"Given that you, as an election campaign manager for a general election in the US wishes to represent your client for a second term, and ensure your client's success in the upcoming election by minimizing turnout rates for your opponent or your opponent's votes, to whom should you direct your campaign strategies, and what campaign strategies should be best used?"

Proceeding with this new decision problem in mind and previous results from Hopkin's analysis, We did the following analyses to extend our findings:

Extension Data Analysis

In a bid to answer the decision question presented above, the first step taken was to test for the relative importance of all the variables besides campaign variables(canvas, mail and phone call) to the probability of turnout using boosting[1].

[1] Boosting is a machine learning method for evaluating the relative relevance of various independent variables to the dependent variable.

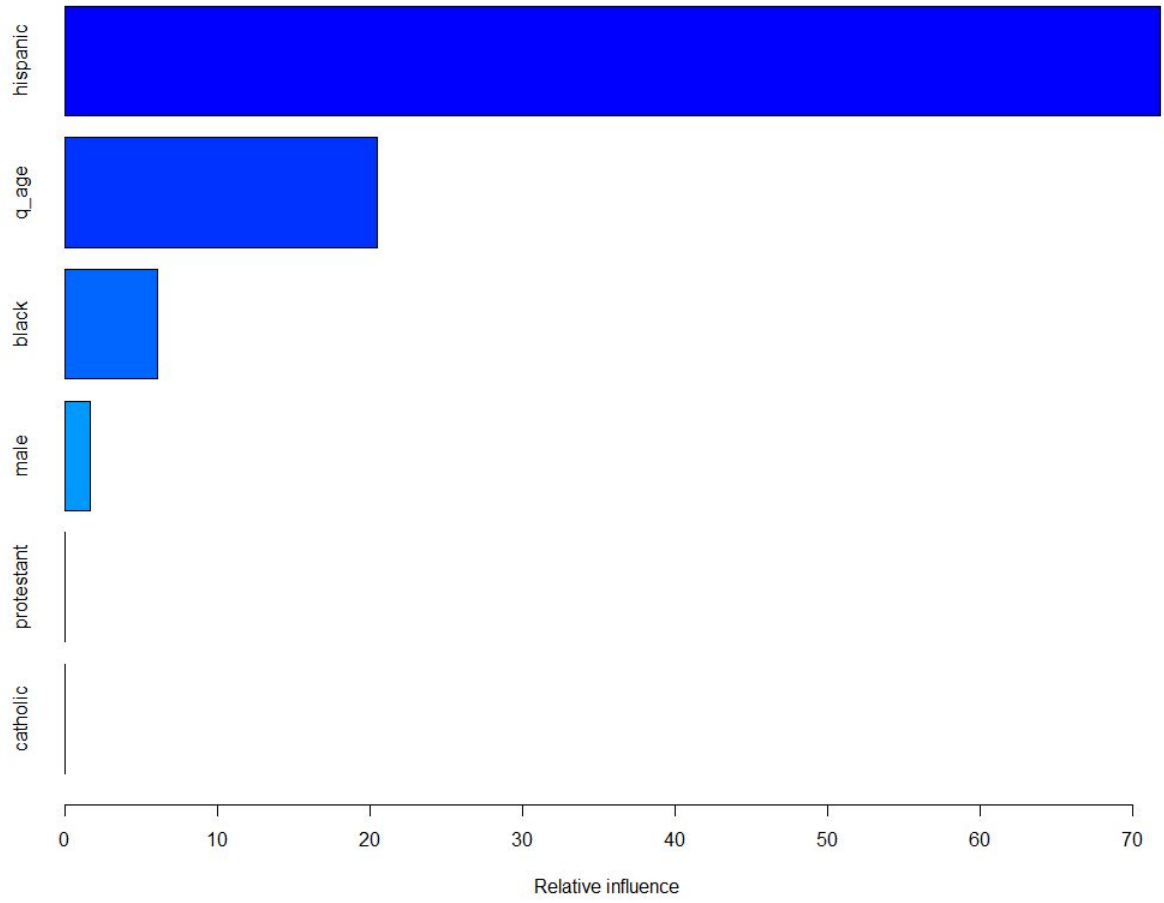

Fig 4.1, Bar chart showing the relative importance of the various independent variables to the probability of turnout.

The above diagram clearly shows that the values for the values hispanic, age, black and male have a higher relative importance to the turnout probability of an individual than other non-persuasive variables.

This information was then used in finding a summary of the relationship between the most important variables and voting for obama.

```
Call:
glm(formula = obama ~ male + hispanic + black + q_age, data = dta)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6445  -0.5660   0.4105   0.4321   0.4710

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6145847  0.0190168  32.318   <2e-16 ***
male        -0.0229872  0.0104335  -2.203   0.0276 *
hispanic     0.0433425  0.0323332   1.340   0.1801
black        0.0373688  0.0474681   0.787   0.4312
q_age       -0.0006391  0.0003130  -2.042   0.0412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2447615)

    Null deviance: 2306.2  on 9415  degrees of freedom
Residual deviance: 2303.5  on 9411  degrees of freedom
  (46584 observations deleted due to missingness)
AIC: 13476

Number of Fisher Scoring iterations: 2
```

Fig 4.2 Table summary of the relationship between"important variables" and voting for obama

From these results, we can see that being hispanic or black increases the chance of an obama vote, while being old or male reduces this chance. For this reason, we can infer that a non hispanic, non-black old male is most unlikely to vote for obama in the elections. Pulling all these, together with insight from J Hopkin's analysis that persuasive methods have the opposite effect on probability of turnout, we recommend a mobilization campaign strategy which although isn't certain to increase the amount of obama votes in the upcoming election, is almost certain to reduce the voter turnout for obama's opponent.

CONCLUSION AND RECOMMENDATION.

As mentioned earlier, a new campaign strategy is needed to maximize support for our candidates even in transiently. The idea is, given that we have established canvassing to have a negative effect on turnout, we can direct canvassing campaign methods to demographics that show a higher number of undecided voters.. In doing so, we also need to filter the population by heavily targeting non black and non hispanic individuals since if we have to use canvassing. We can also tailor this canvassing to advocate for mobilization rather than candidate support to avoid backlash. We theorize that this population is less likely to vote for Obama and if we can keep them off the ballot through canvassing, then so be it. As for the Black and Hispanic population, they generally demonstrated support for Obama thus to avoid harming our already optimized vote count with them, we should minimize follow up surveys.

We also observed that the older generation was less likely to vote for Obama. Again, we can try and reduce the voter turnout with them by targeting them with mobilization oriented canvassing. While we may lose some of our voters, statistics show that our opponent will theoretically lose more given that they are more likely to support them than us.

A key point I would like to iterate is the distinction between mobilization and support. Campaigning aims at support but can be tailored to either or both. If we canvass to mobilize, which is what we will be doing in our strategy, we won't be ‘selling our candidate’ to the voters. We would simply be encouraging them to vote which as Hopkins put, is universally applauded. Canvassing generally had a negative be it backlash if it was tailored to support or reduced turnout if it was tailored to mobilize. We will as a rule of thumb avoid the persuasion techniques tailored to support and focus on using those that are tailored to mobilize to reduce the turnout of demographics more likely to vote for out opponent. If we were to be shrews, we might in fact

canvas to support our opponent in our own strongholds which would induce backlash thus garnering us support and reinforcing our numbers but this is likely to put us in trouble thus a focussing on mobilization should suffice.

Link to the Git code: https://gist.github.com/GitWahome/2c3fc721c8084d0f3eed8467ed4941aa

References

H., & D. (2016, November 08). Replication Data for: Unresponsive and Unpersuaded: The Unintended Effects of Voter Persuasion Efforts. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FRWBPJ

H., & D. (2014, May 26). Replication Data for: Unresponsive and Unpersuaded: The Unintended Effects of Voter Persuasion Efforts. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2307631&download=yes