Indivisible technical evaluation research plan

This data set contains a large sample of voters, voter attributes, and attributes of the areas and states in which they live. The data set presents challenges, particularly because of the possibility that voter attributes or location could influence their treatment condition and canvass result in nonrandom ways. One of the big questions I wrestled with when planning this analysis was whether to try to determine the effect of the program overall (in other words does the program lead to increases in voter turnout as implemented) or try to determine the effect of canvassers successfully speaking with voters. I chose the latter approach, because I felt would offer the must utility to the organization; if you can show that successful voter contacts increase voter turnout, then you can then focus on adjusting the program to maximize the rate of voter contacts.

Secondarily, some information-theoretic model comparison approaches will likely be too computationally cumbersome given the size of the dataset.

My first step in an analysis like this is always to conduct exploratory analysis. In general, I like to first check for visual evidence of an effect of interest in the raw data before proceeding to statistical analysis; if a statistical association is observable too in the raw data, that gives me much more confidence that it is a real effect. I also need to develop an intuition for how predictor terms might be associated with likelihood of voting, the shapes of these relationships, and interrelationships among the predictor terms.

Once exploratory analysis is complete, I plan a three-part analysis; in each step, we will compare the **contacted** group (i.e. those coded as "SPOKEN_TO") to a different control group, each with different expectations of sample bias:

1. **Contacted** vs. **Control** (i.e. those with treatment == 0)
    - The **Control** group may differ from the **Contacted** group based on geography: rural voters (or voters in other areas) may be systematically more likely to not have enough neighbors to allow for control households to be set aside.
2. **Contacted** vs. **Not visited treatment** (i.e. treatment == 1 & canvass_result == "NOT_VISITED")
    - The **Not visited treatment** group may be expected to differ less from the **Contacted** group geographically than comparison 1 above, but these groups could still differ systematically for all kinds of reasons (e.g. something like apparent household wealth influencing both likelihood of a canvasser approaching the house and the residents' likelihood of voting)
3. **Contacted** vs. **Not home**
    - Because canvassers attempted to approach the houses represented in both of these groups, they may be expected to differ less geographically and in terms of external characteristics. However, these groups could differ systematically for other reasons (e.g. residents' willingness to come to the door to talk to a stranger, work-from-home status, etc).

All of these comparisons therefore may involve biased samples with potential biases having unpredictable effects on results. I'm using this approach for three reasons:

1. If all three analyses find consistent effects, results can be considered more robust, given the unpredictable effects of the potential biases outlined above (particularly if these statistical results align with trends in the raw data)
2. Simpler models are easier to describe and interpret, so this approach is preferable to more complicated model specifications if it is tractable
3. Given the size of the data, simpler models will be more efficient

If the results are inconsistent across the three analyses, we can go back and adjust model specification.

For each analysis, I will run a logistic regression generalized linear mixed model (GLMM) with **voted** (1/0) as the binary outcome variable, state ID as a random intercept term, and canvass result, measures of turnout score, partisanship score, age, squared age (to account for the apparent quadratic age effect on voting likelihood), race, urbanicity, and a measure of canvasser enthusiasm as predictor terms.

Then, because statistical significance is a poor measure of the importance and magnitude of a statistical relationship, I will conduct a limited information-theoretic model comparison comparing model fit of models with the **canvass result** term included vs. the same model with that term excluded. This procedure essentially asks whether we do a better job predicting voting behavior using canvassing data than we do without, using a measure of model fit that penalizes overly complex models (AIC).

Finally, if models fit better with the **canvass result** term included, I'll use these models to report the estimated marginal effect of successful canvassing in each case.

With more time, I'd plan to run a more thorough IT model comparison, but that would have to run overnight.

Some thoughts on the prospects of the N2N program after working through the analysis:

- Given that 1) speaking to voters appears to have a consistent positive association with voting rates, but 2) the majority of households in the "treatment" group were coded as not contacted, focus should be given to how to optimize contact success rates for canvassers.
- Scaling may be complicated; I was surprised that the volunteers with the most assignments had lower rates of voter turnout. If these results are to be believed, maybe efforts could focus more on activating first-time canvassers than on encouraging repeat canvassing assignments (although the latter should not be discouraged).
- A follow-up analysis with additional data could be helpful for further optimizing the program: what are the effects of time of day on likelihood of voter contact or (if voter

contact is made) voting behavior? Most people probably don't want to canvas after 5 pm, but would that improve the likelihood of speaking to voters in the household?

- What about the effect of race of the canvasser, or the interaction between canvasser race and voter race, on voter turnout?