

How to Analyze Field Experiments Using Optimal Matching: The Case of the Vote Turnout Experiment

Jake Bowers and Ben Hansen
Political Science and Statistics
University of Michigan

Overview

- Causality and the Importance of Comparison
- Data Example: The 1998 New Haven Vote Turnout Field Experiment (Gerber and Green 2000, Imai forthcoming, Gerber and Green forthcoming)
 - Experiment to Assess the Effectiveness of Phone Calls, Door Knocking, and Mailings on Vote Turnout. (n=29,380)
- Field Experiments and The Problem of Non-Compliance
- Methods of Analyzing “Broken” Experiments: Multiple Regression, Instrumental Variables, Propensity Scores+Matching, Our Hybrid Approach (Matching to strengthen a potentially weak instrument)
- Some results and questions for discussion (this is a work in the early stages of what we hope is progress).

Causality and Random Assignment

- What do we mean when we say “X causes Y”?
 - Treatment Effect for a Person is $Y|X - Y|\text{not}X$
- We most commonly assess causal effects by comparing means within one group to means within another group --- where the two groups are the same except for the treatment.
 - Average Treatment Effect is $E(Y|X) - E(Y|\text{not}X)$
- The reason we like random assignment is because it (in principle) ensures comparability between groups.
- The reason we like field experiments is because they ensure comparable treated and control groups via random assignment + they enhance generalizability.
- Field experiments face threats to comparability because of problems with random assignment (implementation or chance) and because of non-random compliance with treatments.

Some Compliance Problems

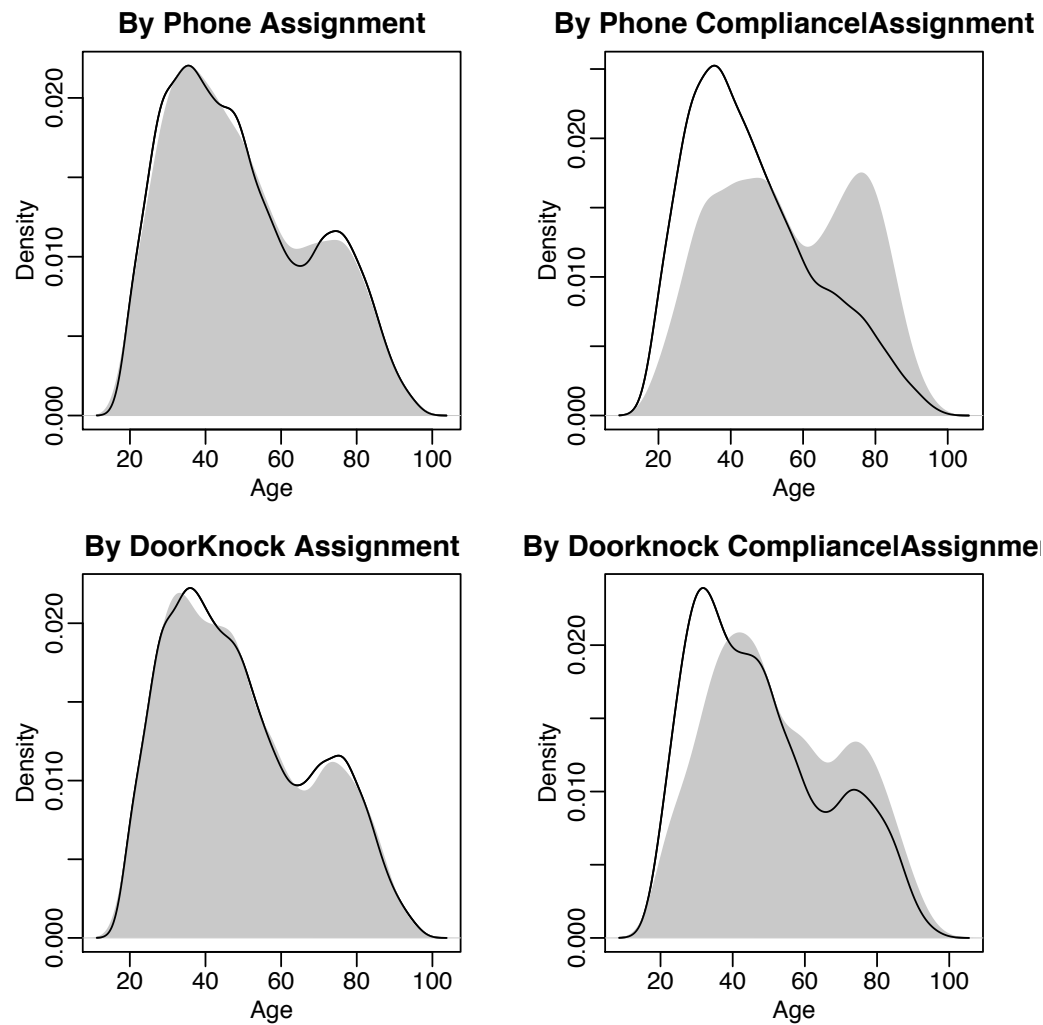


Figure 1: Age Distributions by Assignment versus Compliance

Solutions: Multiple Regression?

- If you know exactly what is causing non-compliance, and the functional form, then you can use multiple regression.
- In this case, although turnout and compliance definitely vary by age, there is probably something more than age leading to these differences.

| | | Treated | | |
|-----------|-----------|---------|---------------|-----------|
| | Age Group | Control | Non-compliers | Compliers |
| Phone | [18,34] | 0.2352 | 0.1957 | 0.4092 |
| | (34,46] | 0.4254 | 0.3720 | 0.5556 |
| | (46,64] | 0.5399 | 0.4966 | 0.6333 |
| | (64,97] | 0.6260 | 0.4936 | 0.6914 |
| DoorKnock | [18,34] | 0.2349 | 0.2223 | 0.3469 |
| | (34,46] | 0.4233 | 0.3933 | 0.5198 |
| | (46,64] | 0.5343 | 0.5302 | 0.6548 |
| | (64,97] | 0.6095 | 0.6271 | 0.7415 |

Table 1: Vote Turnout by Age by Treatment and Compliance

Solutions: Instrumental Variables

- When the assumptions hold, IV can produce a consistent estimate of the overall effect of treatment on the treated (i.e. on those who comply).
- You don't need to know anything about the possible confounders (names or functional form).
- Key assumptions (Angrist, Imbens, Rubin 1996): Treatment is randomly assigned, Treatment only affects Turnout via Compliance.
- Simple to calculate and nice intuition:

$$\text{Turnout}_i = \beta_0 + \beta_1 \text{Comply}_i + \varepsilon_i$$

$$\text{Comply}_i = \alpha_0 + \alpha_1 \text{Treatment}_i + \nu_i$$

- Problems: If random assignment doesn't exist, people say IV won't do.
- We ask: What if random assignment mostly exists?? Even though random assignment is the perfect instrument, can we bolster imperfect random assignment such that we can create slightly weaker instruments? Think in terms of comparability.

Solution: Matching to Produce Comparable Sets

- ❑ Ignore random assignment, pretend that the field experiment is just an observational study. And create comparable sets by stratification or matching.
- ❑ When we want to estimate a treatment effect we want to “control for” possible confounders. “Control for”= “match on”.
- ❑ For example: take Political Donations and Phone Calls. We’d want to control for income, age, and education.
- ❑ If we could match one treated person to one control, where the two people are identical on income, age, and education, then we will have “controlled for” those variables.
- ❑ Advantages: Very intuitive; Automatically alerts you to impossible comparisons; Can use simple t-tests (or nonparametric tests) to compare means; Can easily look at treatment effects by subgroup once you’ve done the match; Model specification is clear --- best model is the one that makes everyone comparable (defined without reference to the dependent variable).
- ❑ Disadvantages: Pairwise matching throws away data (often quite a lot of it); It is hard to find exact matches (and the more variables matched on, the harder it becomes to find exact matches) --- the curse of dimensionality.
- ❑ Solutions: Optimal full matching instead of pairwise matching (new here); Using propensity scores to reduce the dimensions of the variables matched on.

Solutions: Matching and Propensity Scores

- If random assignment has failed, then one approach is to create comparable treated and control groups some other way. One approach is to:
 - Summarize all of the information you have about each person before treatment in a single number (called a “propensity score” because it represents the predicted probability that a given person will be treated based on what we know about them pre-treatment [Age, Ward, Partisanship, Household Size, Voting in 1996]).
 - Match people who have the same propensity score to one another --- one treated person to one (or a few) control person(s).
 - Subtract proportion voting among controls from proportion voting among treated by matched set. Overall treatment effect is average of those differences.
- Key Assumption: The information in the dataset must account for any differences in compliance. Doesn't seem to hold in this case:

| | Propensity Score | Control | Treated | |
|-----------|------------------|---------|---------------|-----------|
| | | | Non-compliers | Compliers |
| Phone | [0.070,0.19] | 0.4730 | 0.3872 | 0.6322 |
| | (0.19,0.21] | 0.4707 | 0.3760 | 0.6059 |
| | (0.21,0.24] | 0.4675 | 0.3962 | 0.6649 |
| | (0.24,0.44] | 0.3949 | 0.3361 | 0.5256 |
| DoorKnock | [0.035,0.18] | 0.4081 | 0.3937 | 0.5833 |
| | (0.18,0.20] | 0.4824 | 0.4944 | 0.6359 |
| | (0.20,0.22] | 0.4976 | 0.4572 | 0.6159 |
| | (0.22,0.49] | 0.4023 | 0.3711 | 0.5356 |

Table 2: Vote Turnout by Propensity Score by Treatment and Compliance

Solution? Match First + IV Later?

- Seems like ordinary propensity score+matching isn't appropriate. Plus, seems a shame to ignore the fact that some random assignment did occur (our preliminary tests show that only pre-treatment imbalance on Ward, everything else looks mostly ok).
- Can we bolster the comparability created by the initial random assignment that was either (1) hurt by implementation problems or (2) luck of the draw?
- Matching is a technique explicitly designed to create comparable sets to enable researchers to assess causal effects.
- IV used to harness the comparability creation power of an even more powerful comparison creator: random assignment.
- Can we use IV after Matching? Especially if the matching is done with an eye toward the substance of the scientific problem?
- Problems with pair-wise matching.
- What we did:
 - Optimal Full Matching --- nearly exactly on age; using propensity score “calipers” to discipline the full matching algorithm. In essence a two dimensional matching.
 - IV estimate (ITT/(proportion comply)) overall, and within age groups.

Results: Overall Effect of Treatment on the Treated

□ Gerber and Green:

- Phone calls decrease turnout by about 2 ± 4 pct points
- Door knocking increases turnout by about 9 ± 5 pct points

□ Us: Overall

- Phone calls decrease turnout by about 5 pct points [not different from zero, as far as we can tell now]
- Door knocking increases turnout by about 14 pct points [different from zero, as far as we can tell now]

Results: Exploring the Mystery of Age and Turnout

Explanations for why some young people vote and others do not must be sought elsewhere.

□ Us: By Age (Highton and Wolfinger 01)

| | Age Group | ITT | TT |
|---------------|-----------|---------|---------|
| Phone | [18,34] | 0.0271 | 0.1193 |
| | (34,46] | −0.0763 | −0.2837 |
| | (46,64] | −0.0022 | −0.0062 |
| | (64,97] | −0.0181 | −0.0336 |
| Door Knocking | [18,34] | 0.0404 | 0.2135 |
| | (34,46] | 0.0273 | 0.0945 |
| | (46,64] | 0.0228 | 0.0752 |
| | (64,97] | 0.0675 | 0.1973 |

Table 3: Treatment Effects by Age Group

- Another benefit of matching is ease of looking at non-linear, non-additive treatment effects. (Would need four instruments for the four effects estimated with 2SLS. Not sure if regression oriented IV estimation can easily look at subgroup treatment effects.)

Take Away / Talking Points

- When compliance or other imbalances hurt random assignment, we would prefer not to ignore the random assignment. We'd like to continue to treat the experiment like an experiment, not an observational study.
- We hope that our approach lays the groundwork for IV, with gentle adjustments for imbalances.
- Even though Angrist et al. talk about IV as The Method to use with random assignment, where random assignment is a (nearly) perfect instrument, economists have talked for years about stronger versus weaker instruments. If some small flaws in random assignment have weakened it as an instrument, we'd like our approach to build it back up to (nearly) full strength.
- Matching is a nice tool --- especially good for teasing out subtleties in causal stories --- useful framework for thinking about data analysis, complements IV.
- For experiments, nice to try to get away from distributional and functional assumptions if possible --- matching is one way to do this.

For the Future

- Optimize size of matched sets for efficiency [some work on this already done, see Ben's JASA paper]
- Optimize some of the fullmatching and associated algorithms especially for large datasets like this one.
- Figure out hypothesis testing within our framework. (some guidance here already from Rosenbaum and randomization inference versus Little and Yau's use of the bootstrap)
- More data on grumpy/un-civic non-compliers?
- Pursue idea that perhaps the youngest voters are also the most susceptible to mobilization attempts --- if you can get them to answer the phone (the door).
- What are the minimal set of assumptions that we have to make to get our method off the ground?