Data task

Sahil Adane

Econometric concerns

- Am I worried about incomplete compliance? In other words, is it possible for a respondent to see either of the facebook ads if I have not assigned them either of the treatments?

  I will assume that the facebook targeting algorithm can be user-profile-specific i.e. I can choose such that only selected user-profiles will see the ad.

- Am I worried about selective data collection in terms of endline survey completion?

  Define a variable called 'missing status'. Missing status is 1 if the person filled the baseline, but not the endline survey. Missing status is 0 if the person filled both the surveys.

  It is possible that the potential outcomes are NOT independent of missing status. For whatever reason, expectation of treated and untreated outcomes could be different based on missing status. For example, people who have a higher and renewed sense of social and individual responsibility during the endline survey are possibly more likely to take the vaccine and fill the survey. In such a case, the groups of people by missing status are not balanced, and potential outcomes (likelihood of getting a vaccine with and without treatment) is not independent of missing status, and thus missing people cannot be dropped without worrying about potential bias.

  For example, I used the first block of code to assign 'missing' status in the endline survey. However, if I had used the second block of code to assign status, the groups by missing status would possibly not be balanced.

```
# missing status assigned randomly
missing_ids <- sample(1:5000, 500, replace = FALSE)
data$missing <- ifelse(data$id %in% missing_ids, 1, 0)
```

```
# missing status assigned based on low endline outcome
missing_ids <- sample(data$id[data$vacc_e == 1], 500, replace = FALSE)
data$missing <- ifelse(data$id %in% missing_ids, 1, 0)
```

- What is the identification assumption?

  Let t_0, t_1 stand for baseline and endline respectively. Y_{i0t_0} stands for control outcome at baseline for individual i. Y_{i1t_1} stands for treatment 1 outcome at endline for individual i and similarly for others.

  Randomization ensures that $Y_{i0t_0}, Y_{i0t_1}, Y_{i1t_0}, Y_{i1t_1}, Y_{i2t_0}, Y_{i2t_1} \perp D_1, D_2 \; \forall i$
  That is $E[Y_{i0t_0}] = E[Y_{i0t_0}|D_1 = 1] = E[Y_{i0t_0}|D_1 = 0]$

  Since the dataset has a time component to it - baseline and endline, this is technically a difference-in-differences regression.

$$Y_{it} = Y_{i0t} + D_1(Y_{i1t} - Y_{i0t}) + D_2(Y_{i2t} - Y_{i0t}) \; \forall t \in \{t_0, t_1\}$$

$$Y_{it} = \alpha_b + \alpha_e I_{t_1} + \beta_1 D_1 + \beta_2 D_2 + \gamma_1 (D_1 * I_{t_1}) + \gamma_2 (D_2 * I_{t_1}) + \epsilon_{it}$$

I_{t_1}, D_1, D_2 are indicator variables for endline, treatment 1 and treatment 2 respectively.

Here α_b estimates the mean vaccine likelihood at baseline. α_e is the average increase in vaccine likelihood from baseline to endline. This captures the time-trend in the control group. β_1 is the mean difference in vaccine likelihood at baseline between treatment group 1 and control group. β_2 is the mean difference in vaccine likelihood at baseline between treatment group 2 and control group. γ_1 is the average increase in vaccine likelihood at endline for treatment group 1 relative to the control group. This estimates the treatment effect for treatment 1. γ_2 is the average increase in vaccine likelihood at endline for treatment group 2 relative to the control group. This estimates the treatment effect for treatment 2.

For the DiD estimate to equal the Average treatment effect on the treated, I need the parallel trend on counter-factual outcomes.

$$E[Y_{i0t_1} - Y_{i0t_0} | D_1 = 1] = E[Y_{i0t_1} - Y_{i0t_0} | D_1 = 0]$$

This is guaranteed the randomization. Then,

$$\begin{aligned} \hat{\gamma_1} &= E[Y_{i1t_1} - Y_{i0t_1} | D_1 = 0] = ATET && \text{by parallel trends on counterfactual outcomes} \\ &= E[Y_{i1t_1} - Y_{i0t_1}] = ATE && \text{by the stronger assumption of randomization} \end{aligned}$$

- Does the result have external validity? Can we expect this ATET for the entire US population?

  For this question, we can collect demographic data for a representative sample of the US population, and see if there is balance across the two groups - people in the experiment and a representative sample of the US population. If so, we can expect a similar ATET for the entire US population.

Analysis

Since I had assigned 'missing status' randomly, the balance test by missing status shows that the two groups - people who completed the endline survey and people who did not complete the endline survey are statistically identical. The age variable does not seem balanced across the two groups. However, I think this is expected because of the multiple comparisons problem - as I test more null hypotheses, the probability of atleast one of the them failing increases. If I test each individual hypotheses more strictly at let's say 0.007 (0.05/7 = 0.007 by the Bonferroni correction), I will get that the groups are balanced. If so, I drop the missing observations without being worried about bias.

Table 1: Balance by missing status

	(1)					
	Mean 1	N1	Mean 2	N2	t-statistic	p-value
Age	49.394	4500	51.564	500	-2.323	0.021*
Education level	2.249	4500	2.240	500	0.313	0.754
Gender	0.504	4500	0.484	500	0.829	0.407
Income category	5.397	4500	5.452	500	-0.513	0.608
Control	0.335	4500	0.314	500	0.973	0.331
Treatment 1	0.331	4500	0.354	500	-1.016	0.310
Treatment 2	0.334	4500	0.332	500	0.070	0.944
Observations	5000					

The next table displays the summary statistics for the three groups.

Table 2: Summary statistics for each group

	(1) Control Mean (SD)	(2) Treatment 1 Mean (SD)	(3) Treatment 2 Mean (SD)
Age	49.338	50.013	48.835
	(20.406)	(20.747)	(20.398)
Education level	2.288	2.231	2.229
	(0.593)	(0.610)	(0.612)
Gender	0.502	0.505	0.503
	(0.500)	(0.500)	(0.500)
Income category	5.364	5.393	5.433
	(2.391)	(2.391)	(2.369)
Observations	1509	1490	1501

I test for balance across treatment group 1 and the control group.

Additionally, I test for balance across treatment group 2 and the control group.

Treatment and control groups seem balanced except for education level. This is either due to multiple comparisons or based on how the missing status was assigned within the treatment and control groups, there is some imbalance. For example, relatively speaking, due to missing status not many control observations got dropped as compared to treatment 1 and treatment 2 observations.

Education is also a categorical variable with levels 1, 2, and 3 and I wonder how balance is affected because of this. On the face of it, a mean-difference of 0.06 at the most (treatment 2 and control) does not seem huge.

Table 3: Balance between control and treatment 1

	(1)					
	Mean 1	N1	Mean 2	N2	t-statistic	p-value
Age	49.338	1509	50.013	1490	-0.898	0.369
Education level	2.288	1509	2.231	1490	2.580	0.010**
Gender	0.502	1509	0.505	1490	-0.167	0.867
Income category	5.364	1509	5.393	1490	-0.330	0.742
Observations	2999					

Table 4: Balance between control and treatment 2

	(1)					
	Mean 1	N1	Mean 2	N2	t-statistic	p-value
Age	49.338	1509	48.835	1501	0.677	0.499
Education level	2.288	1509	2.229	1501	2.659	0.008**
Gender	0.502	1509	0.503	1501	-0.037	0.970
Income category	5.364	1509	5.433	1501	-0.790	0.429
Observations	3010					

Table 5: Average treatment effects

	(1)	(2)
	Standard	With demog. controls
Endline	0.298***	0.298***
	(5.77)	(5.77)
Treatment1	0.0408	0.0344
	(0.79)	(0.67)
Treatment2	0.0286	0.0219
	(0.56)	(0.43)
Treatment1 * Endline	0.302***	0.302***
	(4.23)	(4.24)
Treatment2 * Endline	0.0802	0.0802
	(1.11)	(1.11)
Constant	2.991***	3.155***
	(82.62)	(37.17)
r2	0.0282	0.0310
df_r	8994	8982

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As expected based on how the data was generated, Treatment 1 increases vaccine likelihood on average by 0.3 points whereas treatment 2 has no effect. While the first specification is correct because treatment is randomly assigned, because we have a small sample, I include the demographic variables in the second specification to reduce the standard error.