POLS 7012 Final Exam 2021

Joe Ornstein

Due December 15, 2021 @ 5pm

Abaluck et al. (2021) conducted a large-scale randomized control trial in nearly 600 Bangladeshi villages to assess whether a public health intervention would successfully increase the rate of proper mask-wearing during the COVID-19 pandemic – and if that, in turn, would reduce the rate of infection. You can find a repository of the study's data and Stata code here. For the final exam, we'll replicate some of their findings. Please write and submit an R script (optionally: an RMarkdown document) that conducts the following analyses, knitted to a PDF document:

1. Load two datasets from the repository. First, the `00_common/00_raw/union_baseline_blood_stats.dta` dataset, which reports the percent of residents in each village (`union`) with positive COVID-19 symptoms at baseline – before the intervention began. Second, the surveillance data `surv_data.dta`, which includes 42,207 observations of mask-wearing behavior.
2. What is the advantage of conducting a randomized experiment to answer this research question? If we only had observational data on public health interventions and proper mask-wearing, what are some back door paths you would be concerned about confounding your estimates?
3. Looking at the baseline symptoms data, what is the average rate of COVID-19 symptoms in the treated vs. control villages? Test the hypothesis that the difference in baseline COVID-19 symptom rates between treatment and control villages is zero. Would you say tha the randomization successfully conditioned on pre-treatment rates of infection?
4. Now onto the surveillance data. First we need to clean it up. Create a village-level dataset (i.e. one row for each village) summarizing the percent of residents who were observed properly wearing masks during the study period (`week_gen` 2 through 6).[1]
5. Keep only the villages that have baseline symptoms data, and join the baseline symptoms variable into your new village-level dataset.
6. What was the average rate of proper mask-wearing in the control villages vs. the treatment villages? Plot the relationship (a jitter plot looked nice for me, but do what works for you).
7. Test the hypothesis that the difference in proper mask-wearing rates between treated and control villages is zero.
8. Check out Table 1 of the paper. In the first row, first column, the authors report the coefficient estimate and standard error from a linear model of average mask-wearing, conditioning on four variables: (1) treatment status, (2) pair ID, (3) rate of baseline symptoms, (4) rate of baseline mask wearing. I've looked all over their repository and I can't find the data file with baseline mask wearing. But we have the first three. Estimate a linear model with those three covariates and interpret the results.
9. **Non-Mandatory Bonus Fun**: Why did I make you create a village-level dataset before analyzing the data? Why couldn't you just analyze an individual-level dataset? Using the dataset in

[1] **Hint:** Because each row in `surv` contains multiple observations spread across a bunch of columns (the `mask_a` column), we need to pivot those columns so that each row is a unique observation. Once you've done that, `group_by()` and `summarize()` are your best friends.

`problem-sets/final-2021/potential-outcomes.csv`[2] simulate two sampling distributions for the difference-in-means: one where *individuals* are randomly sampled and assigned treatment, and one where *villages* are randomly sampled and assigned treatment. Compare the standard errors from these two sampling distributions.

[2] This is a fabricated dataset; we know it's fabricated because it includes both the potential outcome under treatment and the potential outcome under control.