Introduction to Causal Inference

Problem Set 2

Professor: Teppei Yamamoto

Due Saturday, July 16 (at beginning of class)

Only the required problems are due on the above date. The optional problems will not directly count toward your grade, though you are encouraged to complete them as your time permits. If you choose not to work on the optional problems, use them for your self-study during the summer vacation.

Required Problems

Problem 1

This problem continues our introduction to R for those of you who are just getting started with the language. *If you chose to solve Problem 2 in the previous problem set, solve this problem.* If you skipped it, this problem is optional for you; however, you may still want to solve it if you are unfamiliar with the `ggplot2` package.

a) Complete Lesson 2 on this website. Again, you should watch each of the video tutorials (Parts 2.1 to 2.7) and complete all the `swirl` quizzes. You should use the `savehistory()` function to save the log of your work and submit it along with the rest of your problem set.

Problem 2

Download from the course website the data from "Monitoring Corruption: Evidence from a Field Experiment in Indonesia" by Benjamin A. Olken. The paper evalutes an effort to reduce corruption in road building projects in Indonesia. One of the treatments, which we focus on here, "sought to enhance participation at accountability meetings, the village-level meetings in which project officials account for how they spent project funds." Before construction began, residents in treated villages were encouraged to attend these meetings.

The outcome of interest to us is `pct.missing`, the difference between what officials claimed they spent on road construction and an independent measure of expenditures. Treatment status is given by `treat.invite`, which takes a value of 1 if the village received the intervention and 0 if it did not. There are four pre-treatment covariates in the data: `head.edu`, the education of the village head; `mosques`, mosques per 1,000 residents; `pct.poor`, the percentage of households below the poverty line; and `total.budget`, the budget for each project.

a) Create a balance table. For each pre-treatment covariate, include comparisons for treated and untreated units in terms of the mean and standard deviation. Report a test, for each covariate, of the hypothesis that the difference in means between treatment conditions is zero.

b) For each covariate, plot its distributions under treatment and control (side by side or overlaying). Include the plots in your writeup.

c) With reference to your table in a) and plots in b), do villages in each condition appear similar in the pre-treatment covariates? Explain the importance of checking balance in a randomized experiment and the result you typically expect to find.

d) Regress treatment on the pre-treatment covariates and report the results. What do you conclude?

e) Using the difference-in-means estimator, estimate the average treatment effect and its standard error.

f) Now estimate the average treatment effect and its standard error using a simple regression of outcomes on treatment. Are the results different from those in e)? Make the changes necessary for them to match exactly, and explain your method.

g) Reestimate the average treatment effect using a regression specification that includes pre-treatment covariates (additively and linearly). Report your estimates of the treatment effect and its standard error. Do you expect them to differ from the difference-in-means estimates, and do they? Explain.

h) (**Optional**) Estimate the ATE for villages with more than half of households below the poverty line, and then do the same for villages with less than half of households below the poverty line. Estimate the standard error of the *difference* in treatment effects and test the null hypothesis that there is no difference between them. What do you conclude?

Optional Problems

Problem A

Consider a field experiment that first uses covariates to divide units into blocks and then employs complete randomization within each block. Units $1, \ldots, N$ were assigned to a single treatment and control under complete randomization in blocks j. J_i is a random variable indicating which of the M blocks unit i belongs to ($J_i \in \{1, \ldots, M\}$). Let the probability of treatment be constant across blocks.

a) Denote the potential outcome under treatment of unit i in block j with Y_{1ij}, and the corresponding potential outcome under control Y_{0ij}. What is the identification assumption that allows us to identify the ATE in this experimental setup?

b) Now, using the identification assumption you wrote down in a), show that the ATE (τ_{ATE}) is identified. (Hint: You'll need to rewrite the ATE to account for the existence of J_i.)

c) Must the following statements be true in the above scenario? Why?

i) $J_i \perp\!\!\!\perp D_i$

ii) $Y_i \perp\!\!\!\perp D_i$

iii) $Y_{di} \perp\!\!\!\perp D_i$

iv) $Y_{di} \perp\!\!\!\perp J_i$

d) Thinking about your own research interests, come up with a scenario where blocking before randomizing treatment might be useful. Say you have a sample of 20 units. Tell us your outcome variable of interest, the treatment you would be randomly assigning, and the variable(s) on which you would like to block. Briefly explain the benefits to blocking in this design. Are there any challenges to blocking?

e) You explain your research design to your friend who says "You don't need to block on those variables. You can just control for those covariates in a regression after you randomize, run the experiment and collect your data. Covariate adjustment is just as good as blocking!" Is he right or wrong? What concerns might you have with your friend's approach?

Problem B

In some research contexts, assigning treatment at the unit level is impossible. Researchers may instead assign clusters of units to treatment. Here, we use simulation to examine the precision of estimates under this design. Consider the following scenario. We have units i in $1, \ldots, N$ grouped in clusters j in $1, \ldots, M$, where each cluster j has size N/M. Treatment $D_i \in \{0, 1\}$ is assigned via complete randomization at the level of clusters: in $M/2$ clusters, all units are treated while in the remaining clusters, no units take treatment.

a) Create a dataset that matches this description in `R`, as follows. Create a vector of j_i for $N = 100$ so that there are 10 clusters, and randomly assign five to treatment and the rest to control. Induce intracluster correlation by drawing group means μ_{dj} for each cluster: $\mu_{0j} \sim \mathcal{U}(-1, 1)$ and $\mu_{1j} \sim \mathcal{U}(-0.5, 1.5)$. Then draw values for y_{0i} from $\mathcal{N}(\mu_{0j}, 1)$ and for y_{1i} from $\mathcal{N}(\mu_{1j}, 1)$. Using D_i, create a vector of realized outcomes y_i. Now, wrap your work in a function and for each of 1,000 iterations, estimate the average treatment effect, forming a sampling distribution for the ATE. Repeat all of these steps for $M = 20$ and $M = 50$, and plot the three resulting sampling distributions.

b) Explain what happens to the distribution of estimates as M increases, holding constant N.

c) Repeat your simulation for $N = 200$ and $N = 400$ with $M = 10$ and plot the results. Given 100 units in 10 clusters, are there larger gains to precision from doubling the sample size or the number of clusters?

d) Is this result an artifact of our scenario, or does it generalize? Explain with reference to Equations 3.4 and 3.22 from Gerber and Green (2012) for the standard error of the ATE under complete random assignment without clustering and under cluster random assignment.