PS 207: Problem Set 2

Professor Matto Mildenberger

Due Tuesday 30 April before class

Instructions:

- Responses should be typeset in LaTeX, or through Rstudio/Rmarkdown or similar if possible. If you have no background in these tools, that is okay for now but you should start learning it.
- As always, it is HIGHLY recommended you work in groups on this problem set. Some of the problems are meant to stretch your coding skills (e.g. Problem 2a) and are designed to be solved collaboratively.

Problem 1: Analyzing Experimental Data

For this problem we will use data from Benjamin A. Olken. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy.* 115: 300-249. The paper and the data set are available on the course website (`OlkenData` and `OlkenPaper`).
The objective of this experiment was to evaluate two interventions aimed at reducing corruption in road building projects in Indonesian villages. One treatment was audits by engineers; the other was encouraging community participation in monitoring. This problem focuses on the latter intervention, which consisted of inviting villagers to public meetings where project officials accounted for budget expenditures. The main dependent variable is *pct_ missing*, a measure of the difference between what the villages claimed they spent on road construction and an independent estimate of what the villages actually spent. Treatment status is indicated by the dummy variable *treat_ invite*, which takes a value of 1 if the village received the intervention and 0 if it did not.
The variables in the data set are:

- *pct_ missing*: Percent expenditures missing
- *treat_ invite*: Treatment assignment
- *head_ edu*: Village head education
- *mosques*: Mosques per 1,000
- *pct_ poor*: Percent of households below the poverty line
- *total_ budget*: Total budget (Rp. million) (determined prior to intervention)

a) Estimate the average treatment effect in this new dataset, using the difference in means estimator.

b) Ignore concerns about sampling with replacement and just thinking of your $Y_i|D_i = 1$ and $Y_i|D_i = 0$ as being sampled independently of each other. Under these conditions, derive a simple estimator for the standard error of the above difference-in-means.

c) Use the data to estimate the standard error you derived in (b).

d) Check the covariate balance in this dataset on all covariates (all variables that are not the treatment assignment or the outcome vectors). Decide on a test statistic for testing the balance of each covariate and report these in a table. Decide on a few sensible balance statistics and report them in a table. How do the treatment and control group differ? [Hint: Consider using the Matching or ebal packages here to automate this task. I will leave it as a problem set exercise for you to learn how to use these packages and find the relevant functions within one or the other function)

e) Now use regression to estimate the $SATE$ (sample average treatment effect). Is this estimate different from the difference-in-means estimate? Be sure to use an appropriate choice of standard error.

f) Using your answer for (b), conduct a t-test of the null hypothesis that $SATE = 0$. You may use a normal approximation for the cutoff value.

h) Is the standard error of the OLS estimate different than the standard error of the difference-in-means estimate? Why or why not?

i) Re-estimate $SATE$ using three additional regression models: one in which you include all pre-treatment covariates as additional linear predictors, another in which you include arbitrary functions of the covariates (polynomials, logs, interactions, etc.) as additional linear predictors, and a third in which you interact the treatment variable with a demeaned covariate $(X_i - \bar{X})$. Report the treatment effect estimates and their robust standard errors. How do these results vary across the regressions?

Question 2. Randomization Inference

a) What is the sharp-null, and how does it compare to the null hypothesis we tested in the previous question?

b) Why is the sharp null a convenient choice? I.e., what special property of the sharp null allows us to obtain a distribution of outcomes for a test statistic under this null?

c) Write your own function in `R` that takes as arguments a vector of outcomes, Y, and the original treatment assignment vector, D, and produces (i) a plot showing the distribution of the difference in means statistic under the sharp-null, (ii) a vertical line representing the observed difference in means relative to this distribution, (iii) a p-value for the difference in means statistic against the sharp-null. (Since you do not know the length of Y or D yet, rather than trying every permutation exhaustively, you may simply try a large number of permutations.) [Hints: This can be done using ONLY the following R commands - function(), length(), mean(), rep(), sample(), plot(), abline(), sum(), return().]

d) Apply your function to the data from the Olken experiment in the previous section. How do your results compare to the results under the t-test or regression? Which do you trust, and how do you interpret any differences?