# Political Science 207
# Problem Set 2

Professor: Daniel Masterson

Due Tuesday April 20th at 5:00 p.m.

Submit your completed assignment as a single PDF or HTML file (including your write-up and all R code) on GauchoSpace under the assignment for Week 3. I strongly recommend that you write your problem sets in R Markdown. (I recommend this tutorial.)

Make sure you follow good coding style (see tutorial in link), and show your work for problems that require calculations.

You are encouraged to work in groups, but you should write up the problem set alone, and you should note at the top of your problem set who you worked with.

## Problem 1

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: 200-249 (`olken_data.csv`). 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 the dataset, using the difference in means estimator. (Make sure to first remove any units with missing value(s) for one or more variables.)

b) On what characteristics (i.e., covariates) do the treatment and control group differ, if any?

c) Now use regression to estimate the $ATE$ (average treatment effect). Use heteroskedastic-consistent standard errors. Is this estimate different from the difference-in-means estimate?

d) Re-estimate the $ATE$ using three additional regression models: (1) one in which you include all pre-treatment covariates as additional linear predictors, (2) another in which you include arbitrary functions of the covariates (polynomials, logs, interactions, etc.) as additional linear predictors, and (3) a third in which you include demeaned versions of the covariates ($X_i - \bar{X}$) as well as the interactions between each of them and the treatment. Report the treatment effect estimates and their heteroskedastic-consistent standard errors. How do these results vary across the regressions?

e) What can you conclude about the effect of community monitoring on corruption, given your analysis above?

# Problem 2

Are the following statements true or false? Justify your answers for credit.

a) A key identification assumption in a randomized experiment is that there is no interference between units (the SUTVA assumption). To make one's inference robust to the violation of that assumption within clusters of units, one must correct standard errors using methods such as cluster-robust standard errors.

b) The potential outcomes framework assumes that treatment assignment of unit $i$ does not affect the potential outcomes of unit $j$.

c) The following identity generally holds: $ATE = \frac{1}{2}ATT + \frac{1}{2}ATC$ where $ATE$ is the average treatment effect, $ATT$ is the average treatment effect for the treated, and $ATC$ is the average treatment effect for the controls.

# Problem 3

You are on a team of researchers who want to estimate the effect of having completed college in the past ($C_i = 1$) on an individuals' current earnings ($Y_i$). Some individuals may have taken up a white collar job after finishing college ($W_i = 1$), but before you measure earnings. Assume that whether a subject completes college is randomly assigned. Your fellow researcher says that you need to control for the occupation of a person in your model. To do this they write the difference in means of potential outcomes for all individuals who have a white-collar job as follows:

$$E(Y_{1i}|W_{1i} = 1, C_i = 1) - E(Y_{0i}|W_{0i} = 1, C_i = 0)$$

a) Why does $W$ have two subscripts? What do they represent, in terms of this study?

b) Random assignment of $C$ allows you to re-write the above expression as:

$$E(Y_{1i}|W_{1i} = 1) - E(Y_{0i}|W_{0i} = 1)$$

Is this difference unbiased? Why or why not?