NYU Wagner
Advanced Empirical Methods
Rajeev Dehejia

Problem Set 1: Randomized Evaluations

The goal of this problem set is to analyze data from a randomized impact evaluation. We will be using data from a paper by Thornton (AER 2008) titled, "The Demand and Impact of Learning HIV Status". The questions on this problem set will lead you through a series of exercises that are standard practice when analyzing data from a randomized evaluation. A copy of the paper is on our web directory. You should at least read pages 1829-1839

Please submit a do file, log file, and a write-up of no more than 3 pages for the problem set. Written responses should be *brief*.

The data are on our web directory.

Preliminaries:

Thornton (2008) examines whether varying the cost of HIV testing can increase the number of people who get their test results. There are two interventions: 1) cash payments to individuals who receive their HIV test results; and 2) the distance a person needs to travel to obtain their HIV test results. Both interventions are randomly assigned on an individual level.

Key Variables:
AnyCash = 1 if randomly assigned any cash incentive to obtain HIV test result
under = 1 if randomly assigned distance to get test result is under 1.5 km
Cash Amount = the randomly assigned cash incentives (amounts) to obtain HIV test results (where 0 corresponds to the control group)

Going forward, there are two treatments: those who were assigned a cash incentive and those who have to travel less than 1.5 km to get their test results. Depending on the case we may analyze these individually or separately.

Questions

Part I: Summary Statistics

The first step is to look at summary statistics of your sample. This will tell you the sample that you are analyzing. We will also see if there are differences between the treatment and control group.

1. Present summary statistics for the study sample. What is the average age? What percentage of males are in the study? What percentage of people are infected with HIV?
2. Present summary statistics for those in the control and treatment group, separately for the any treatment and the distance treatment. Are there major differences in any of the variables (i.e. age, education, HIV rates)?
3. Do a test to see whether *differences in age, HIV rates, and marriage are statistically different between the treatment and control group; consider separately the treatment and control groups defined by any and by distance*. Do you see any differences? Are you concerned by any of the differences? How could they affect the analysis?

Part II: Analysis using graphs
We can create simple graphs that can help us see the effects of the treatment.

4. Generate a bar graph, where the X-axis represents the *any* control and treatment group, and the y-axis is the percentage in each group that learn their HIV status. Let the treatment group in this question be anyone who receives a cash incentive to obtain their HIV test results.
5. Now, generate the same bar graph, but this time varying the amount of cash that people receive in treatment (use the "Ti" variable for the x-axis).

Part III: Analysis using linear regression
Using OLS regression analysis is one of the most common tools used to estimate the effect of a treatment or randomly assigned intervention.

6. a. Consider the following regression output from the data you are using:

```
      Source |       SS           df       MS      Number of obs   =     2,317
-------------+----------------------------------   F(4, 2312)      =     30.58
       Model |  19.3844475         4  4.84611187   Prob > F        =    0.0000
    Residual |  366.412704     2,312  .158483003   R-squared       =    0.0502
-------------+----------------------------------   Adj R-squared   =    0.0486
       Total |  385.797151     2,316  .166579081   Root MSE        =     .3981

------------------------------------------------------------------------------
 usecondom04 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    educ2004 |   .0080766   .0027522     2.93   0.003     .0026795    .0134736
        male |   .1379404   .0175444     7.86   0.000     .1035361    .1723447
         age |  -.0037836   .0007116    -5.32   0.000    -.0051791   -.0023882
    land2004 |  -.0304496   .0210122    -1.45   0.147    -.0716543     .010755
       _cons |   .2808191   .0363461     7.73   0.000     .2095447    .3520935
------------------------------------------------------------------------------
```

Which variable is being regressed on which variables? Provide a statistical and substantive interpretation of the five estimated regression coefficients. Also provide an interpretation of other output generated by the regression, included SS, MS, Root MSE, R-squared, Adjusted R-squared, F(4,2312), Prob > F.

b. Run the following OLS regression, where getting your HIV test result is the dependent variable, and receiving a cash incentive as your covariate.

$$\textit{Got Test Result} = a + b\textit{AnyCash Incentive} + u$$

What is your estimate of b? Is it statistically significant? If so, at what level? What happens when you include additional control variables (age, male, education, marriage)? Does your estimate of *b* change? Does this suggest that there is covert or overt bias?

7. Conduct the same analysis as in question 6 but using a group means comparison. What is your estimate of the treatment effect? How does this differ from queston 6? What assumption holds?

8. Now run a similar regression, but this time replace "Any Cash Incentive" with "Cash Amount". What is your estimate of b? Is it statistically significant? What happens

when you include additional control variables (age, male, education, marriage)? Does your estimate of b change?

$$\textit{Got Test Result} = a + b\textit{Cash Amount} + u$$

9. Now interpret your findings. Based on your estimates from Q6, what can you say about the effect of offering cash incentives on people learning their HIV status? Would you say that this is a big or small effect? Now look at your estimates from Q8. Does a doubling of the cash incentive from $1 to $2 have a big effect on people's willingness to get their HIV test results? Does this surprise you?

Part IV: Conditional (Heterogeneous) Treatment Effects

We might be interested in whether the treatment has different effects for sub-populations. For example does giving cash incentives have a different effect for men and women?

10. Create an interaction term which interacts the treatment (any) with the gender variable (male). Run the following regression:

$$\textit{Got Test Result} = a + b\textit{AnyCash Incentive} + c\textit{Male} + d\textit{Any Cash Incentive X Male} + u$$

What is your estimate of *d*? Is it statistically significant? How do you interpret this result? How does the interpretation of d differ from the interpretation is *c*?

11. Create an interaction term that interacts treatment (any) with education (educ2004). Run the following regression:

$$\textit{Got Test Result} = a + b\textit{AnyCash Incentive} + c\textit{Education} + d\textit{Any Cash Incentive} \times \textit{Education} + u$$

What is your estimate of d? Is it statistically significant? How do you interpret this result?

Part V: Policy Implications

12. Based on your findings from Part III, if the goal of the government were to increase the number of people who know their HIV status, what type of policy would you recommend?
13. Based on your findings from Part IV, are there certain groups that cash incentives should target?

Part VI: A Random Sub-Sample

14. Now suppose that you have to choose a random subsample from the population (which is analogous to randomly choosing units for treatment). You should randomly choose 1,000 people from your entire sample.
15. Now rerun the same regression as in Q6. Is your estimate of *b* different? Why?

Part VII: Choosing Sample Size

16. Suppose you were running an experiment in India looking at the effect a cash incentive on condom purchases. You want to use this data to get a sense of how big an experiment you will need. Suppose you want to detect a difference between the treatment (cash incentive) and control (no incentive) groups of 1 condom purchase. Based on Thorton's data, how large a sample would you need? For convenience assume that the *any* variable is the relevant treatment. Assume size of 0.95 for the

test, and try power levels of 0.8 and 0.9. Assume treatment and control groups of equal size.

17. Now suppose that your data will be clustered (e.g., individuals in villages). You need to adjust for the fact that what goes on within a classroom isn't independent. You can do this using the sampclus correction after you run each of the previous commands. But before doing that you'll need to estimate the intraclass correlation. How correlated are observations within each cluster? First use the loneway command to estimate the intraclass correlation of condom purchases by village. Then assuming 40 observations per village, calculate how many villages you will need to detect a one condom purchased difference with powers of 0.8 and 0.9 respectively.

Part VIII: Fisher Randomization Test (bonus)

18. In this question I ask you to implement a Fisher randomization test over the variables *got* and *any*. First compute the simple difference in *got* between any=1 and any=0. Then using the sharp null of no effect, simulate the probability of observing a mean difference greater than the difference we in fact observe in the data. What do you get? Then simulate the probability of observing a difference greater in absolute value than 0.05 and 0.01. Hint: be sure to exclude (or drop) any observations where *any* is missing (keep just the 0 and 1 observations).