Harris School of Public Policy
PPHA 346 Program Evaluation
Winter 2018

Assignment 1

Due date: Jan. 24, 2018, by 11:59pm via Canvas. Your answers must be typed; submit a .doc or .pdf document.

Format: please submit your answers in the following format. You may copy-paste Stata outputs by simply select, right click, copy as picture.

1. Question

Written explanation

```
stata code
```

Stata output

For example:

1. What is the mean of X?

The mean of X is 5.33.

```
. mean X

Mean estimation                 Number of obs    =     21646

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
           X |   5.331285   .0067868       5.317983   5.344588
--------------------------------------------------------------
```

This will make you show every line of your code throughout your answers, so you do not have to also submit a do-file. Some questions will only need code and no outputs, so in those cases simply present the code you typed.

I. Manski Bounds

In this exercise, you will perform calculations similar to those in slides 23-28 from lecture 1. Note that in the slides, the outcome variable is "being a recidivist," so a dummy $Y_i = 1$ if the individual is a recidivist and $Y_i = 0$ if she is not. The treatment variable is "being assigned to juvenile detention," also a dummy $D_i = 1$ if the individual is assigned and $D_i = 0$ if she is not. Given that juvenile detention is compulsory if you are assigned to it, assume perfect compliance; therefore, "assigned to juvenile detention" = "juvenile detention."

You will be working with prison population data from Iowa in 2013. Download the .csv dataset from the link below and import it to Stata.

https://catalog.data.gov/dataset/3-year-recidivism-for-offenders-released-from-prison

Your outcome variable will also be "being a recidivist." Use the variable `recidivismreturntoprison`, which indicates "Yes" if the individual is a recidivist and "No" if otherwise. Your treatment variable will also be a dummy indicating if the person participated in a program implemented by the prison system or not. Use the variable `partoftargetpopulation`, which says "Yes" if the person is part of the target population of the program and "No" if otherwise. Also assume perfect compliance, so "part of target population" = "participant."

1. Use `recidivismreturntoprison` and `partoftargetpopulation` to generate new dummies with numeric values of 0 and 1. Call your outcome variable Y and your treatment variable D. After you create them, you should verify they match—Y should show a 1 if `recidivismreturntoprison` shows a "Yes," etc. Some of you might need to use the `encode` command (type `help encode` on Stata to learn more).

2. Using these two new dummies, complete the following table of probabilities rounding to two decimals.

Pr(Participant)	
Pr(Not participant)	
E(Recidivist)	
E(Recidivist | Participant)	
E(Recidivist | Not participant)	
N	

(You're not required to use Stata for performing calculations in questions 3-9).

3. Using the naïve estimator, what is the average effect of participating in the program?

4. Interpret this result. What does the naïve estimator suggest about the impact of Iowa's prison system program on recidivism? What are the underlying assumptions?

5. Start bounding the impacts by calculating $E(Y_{1i})$. What is the missing counterfactual here? Interpret in a sentence what this term in your equation means.

6. Calculate $E(Y_{1i})^{UB}$ and $E(Y_{1i})^{LB}$.
7. Now calculate $E(Y_{0i})$. What is the missing counterfactual? Interpret in a sentence what this term in your equation means.

8. Calculate $E(Y_{0i})^{UB}$ and $E(Y_{0i})^{LB}$.

9. Lastly, calculate $(\Delta^{ATE})^{UB}$ and $(\Delta^{ATE})^{LB}$. What are the bounds of the average treatment effect? Is the naïve estimator within this range? Interpret in a sentence what the upper and lower bounds imply about the impact of Iowa's prison system program on recidivism. What are the underlying assumptions of Manski's bounds? What are some downsides of bounding?

II. The Naïve Estimator

For this exercise, you will benefit from reading the materials posted for the first TA session. Open `ENIGH.dta`. This is a subset of variables from the 2016 Mexican National Survey for Household Income and Expenditures (ENIGH). The variables are as follows:

`job_earnings` is quarterly household earnings from jobs (excludes all non-job income) in Mexican pesos. This is the main source of income for most households.
`head_gender` is the sex of the household head, so defined by household members. Takes two values:
1 male
2 female
`head_age` is the age of the household head.
`head_schooling` is the education of the household head. Takes the values:
1 non-response
2 no education
3 pre-elementary
4 elementary school
5 middle school
6 high school
7 teaching school (a degree for becoming a teacher)
8 vocational/technical school
9 college
10 master's degree
11 doctorate degree
`tot_members` is the total number of members living in the household
`men` is the number of men living in the household
`women` is the number of women living in the household
`age12_64` is the number of people living in the household who are 12 to 64 years old
`workers` is the number of people living in the household who have a job
`loc_size` is the population size of the locality where the household is. Takes the values:
1 Population size 100,000 or larger
2 Population size 15,000-99,000
3 Population size 2,500-14,999
4 Population size less than 2,500

1. You will be testing the impact of having a male household head on job earnings. The outcome variable is therefore `job_earnings`, and the treatment variable here is "having a male-headed household." Create a new dummy variable called `head_male` equal to 1 if the household is headed by a man and 0 if headed by a woman.

    a. What share of households have a male head? A female head?

b. Calculate the naïve estimator by comparing two conditional means, and then by running a bivariate regression. Interpret this result and state what the underlying assumptions are.

2. Assume that, if there are any selection problems, these are perfectly and entirely captured by the remaining variables included in your dataset (so there could only be selection on observables). Now, you will test the assumptions of the naïve estimator by looking at differences in these variables between the households led by males and the households led by females.

   a. Use the information in the variable `head_schooling` to calculate what share of male-headed households do or don't have a college-educated head. Do the same for female-headed households. You might benefit from creating a dummy `head_college` which equals 1 if the head went to college, and 0 if he/she didn't. Take categories 2-8 as "no college" and categories 9-11 as "college." You may use the `inrange` function.

   b. Choose two of the following six variables: `head_age, tot_members, men, women, age12_64, workers`. For one of them, graph a boxplot of its distribution conditional on having a male or female household head. For the second one, plot two overlaid histograms to compare its distribution conditional on having a male or female household head. Properly label both graphs. Write a few sentences on the similarities or differences you find.

   c. Look into urban vs. rural figures. Define a locality as rural if it has a population of less than 2,500 and as urban if otherwise, using `loc_size`. What share of households are in rural areas and in urban areas? Is the percentage of female-headed households higher in urban or in rural areas?

   d. Now, formally test if the treated households (male head) are systematically different from untreated households (female head) by using simple bivariate regressions. Test this for variables `head_age`, `head_college`, `tot_members`, `men`, `women`, `age12_64`, `workers`, and `rural`. Interpret these results. On average, do households led by males look similar to households led by women?

   e. Is the naïve estimator biased here? Why or why not?

   f. Recalling the assumption we made at the beginning of question 2, specify a multivariate regression model that gives you an unbiased average treatment effect. Use `head_male` instead of `head_gender`; `head_college` instead of `head_schooling`; and `rural` instead of `loc_size` for interpretation purposes. Also, instead of `men` or `women` (you cannot use both due to multicollinearity), rather use the share of men out of total household members (`men / tot_members`). Run your multivariate regression and

show the results. (It is okay to find that one covariate is not significant).

g. What average treatment effect do you get? Is this similar to your naïve estimator? If not, why do they differ and by how much? From your perspective, are there any interesting findings regarding the other variables?

3. Now we travel to the remote country of Randomburg, and collect practically the same variables as Mexico's ENIGH. You can find them on `Randomburg.dta`.

   a. What is the naïve estimator of the impact of `head_male` on `job_earnings`?

   b. Assume here, too, that if there were any selection problems, they would be perfectly and entirely captured by the variables included in our dataset. Now, test if there are differences between those households with male heads and with female heads in Randomburg. Jump right into the formal tests using regressions like in step (d) from question 2. Test this for the variables `head_age`, `tot_members`, `age12_64`, `workers`, `head_college`, and `rural`. On average, do households led by men look similar to households led by women?

   c. Assume here, too, that if there were any selection problems, they would be perfectly and entirely captured by the variables included in our dataset. Is the naïve estimator biased here? Why or why not?

   d. Run a multivariate regression of `job_earnings` on `head_male` where you control for all covariates in your dataset and show the results.

   e. What treatment effect do you get? Is this similar to the naïve estimator? Why? What do you conclude?