Problem Set 1
Econ 672
June 30, 2022

1) From the dataset "calculate_ate_sdo.dta", we will need to calculate the SDO, ATE, ATT, and ATU from the data set. We assume that we know the potential outcomes Y0 and Y1 (in reality we observe only Y0 or Y1 but never both at the same time) and we have treatment status d and an ID variable called id.
    a. Calculate the SDO and ATE. How bias is the SDO from the ATE? Is it upward or downward biased?
    b. Calculate ATT and ATU, does $ATE = (\pi) * ATT + (1 - \pi) * ATU$ where $\pi = \frac{N_T}{N}$
    c. Calculate the selection bias?
    d. Calculate the heterogeneity treatment bias?
    e. Does $ATE = SDO + SelectionBias + HetrogeneityTreatmentBias$?

2) A small set of the 2017 Q4 Public Use PIRL data from DOL Employment and Training Administration can be found on ELMS called "pirl_small.dta" (it is too large for GitHub). This file has been limited to all individuals who recorded employment 4 quarters after exiting the program and the variables have been renamed for more intuitive understanding. You need to find the causal effect of training on employment and wages 4 quarters after program exit. Some individuals got training, but most did not. Individuals are endogenously sorted into training based upon characteristics, especially individuals who have significant barriers to employment. Utilize a propensity score matching methodology to answer the following questions. Real world data come with a data dictionary. The PIRL data dictionary can be found here: https://www.dol.gov/sites/dolgov/files/eta/performance/pdfs/PY2017/WIOA%20Performance%20Records%20Public%20Use%20File%20Record%20Layout%20PY2017Q4.pdf or on ELMS and GitHub. (Hint 1: the trimming rule of thumb of [.1,.9] might not be appropriate and try trimming from a lower bound below 0.1). (Hint 2: ***do not to run teffects before trimming and common support*** or else there is a good chance Stata will freeze up).
    a. Provide common support tests from two different methods discussed in class.
    b. What level did you trim your data for common support? Why?
    c. What is the estimated ATE on employment with Inverse Probability Weights? What is the estimated ATE on employment with K-nearest neighbors PSM? How does this compare to the SDO?
    d. What is the estimated ATE on wages with Inverse Probability Weights? What is the estimated ATE on wages with K-nearest neighbor PSM? How does this compare to the SDO?
3) Use the dataset "social_insure.dta" with an instrumental variable methodology for the following problem. Cai, De Janvry, and Sadoulet (2015) examine the decision about buying insurance again weather events and they were interested in whether information travels through social networks. Begining on page 98 in the "Social Networks and the Decision to Insure", the authors investigate how much does what your friends learn about insurance affect your own takeup rate. $Takeup_{ij} = \alpha + \delta TakeupRateFriends_j + X'\beta + \varepsilon_{ij}$ is the equation used where i is for individuals in region j. The problem is that the TakeupRateFriendsj is endogenous to Takeup for

an individual i in region j buying insurance. In order to get around the endogeneity, the authors utilize an instrument called "default" in the first stage and compare that to see how farmers were affected in the second stage in rural China.

We want to identify the effect of Friends Purchase Behavior on Your Purchase Behavior. The outcome of interest is "takeup_survey" (did not buy insurance=0; buy insurance=1). The endogenous treatment of interest is "pre_takeup_rate" (continuous) for the friend's purchase behavior. The effect has backdoors, such as preferences for insurance may be higher or lower by region or you have may have similar preferences. "default" is an instrument for friend's purchase behavior. This is a binary indicator for whether your friends were **randomly assigned** to a "default buy" informational session, where attendees were assigned to buy insurance by default with a preference not a buy ("default==1"). Or "default no buy" session, where attendees were assigned to not buy insurance by default with a preference to buy insurance ("default"==0). There are additional covariates that can be used in the first stage. (Hint, if you want to set fixed effects dummies for a region that is a string, use the encode village, gen(villageid)).

a. Estimate the biased OLS (linear probability model) without the instrument. Estimate the Two-stage least squares with the instrument.
b. What is the F-statistic in the first-stage? Is this a good instrument? Why is it a good instrument? Looking in Cunningham (2021), what kind of popular instrument would this be?
c. What is the estimated local average treatment effect (LATE) for farmer compliers from the second stage?
d. If we use OLS to assess the endogenous treatment "pre_takeup_rate" on takeup_survey, what is the estimated coefficient? How does it compare to the estimated ATE in section c? Is it upward or downward biased?
e. We need to test assumptions. Does this instrument violation the non-zero first stage assumption? Does this instrument violate the monotonicity assumption?
f. Do we need to cluster standard errors by any group? If so, how does it affect the standard errors around the LATE?