ECON3096 - Causal Inference
Problem Set 1 (Lectures 1-3)
Due Date: 10 October before class (15:30 pm)

1. Consider the following three causal questions:

   - Many firms, particularly in southern European countries, are small, and owned and run by families. Are family owned firms growing more slowly than firms with a dispersed ownership?
   - What is the effect of taking a job in a finance position compared to a human resources position on the salary earned by an economics graduate?
   - What is the effect of mortgage interest rates on the number of new housing starts?

   For each of these questions answer the following:

   (a) What is the outcome variable and what is the treatment?
   (b) Define the counterfactual outcomes $Y_{01}$and Y_{1i}.
   (c) What plausible causal channel(s) runs directly from the treatment to the outcome?
   (d) What are possible sources of selection bias in the raw comparison of outcomes by treatment status? Which way would you expect the bias to go and why?

2. Coming back to this example we had dicussed in class:

Suppose the population for variable Y_i consists of

unit	1	2	3	4	5	6	7	8	9	10
Y_i	4	2	5	5	3	1	2	5	4	5

Population mean: 3.6

Samples:

Sample 1

unit	1	2	4	8
Y_i	4	2	5	5

Sample avg: 4

Sample 2

unit	2	5	9
Y_i	2	3	4

Sample avg: 3

   (a) What are the variance and standard deviation of the population mean $\overline{Y_i}$?
   (b) What is the standard error of the means of Samples 1 and Sample 2?
   (c) Suppose now we want to examine if the sample means of Sample 1 and Sample 2 are significant different from each other, what kind of test we shall use?

(d) What is the exact statisics number and what does it tell us about the difference?

3. Suppose we are interested in find out whether increasing the availability of computer for student could help to improve their test scores. We start from two variables: the number of computers per student and test scores for 5th graders, `comp_stu` and `testscr`.

   Download the data set `caschool.dta` and read it into **R**. You could use the code in "ps2.R" to read in the dta dataset to **R**. It contains observatiosn on 420 California school districts in 1999 on 14 variables. Description on these variables is as the following:

   - district: District code.
   - school: School name.
   - county: County name.
   - gr_span: Grade span of district.
   - enrl_tot: Total enrollment.
   - teachers: Number of teachers.
   - calw_pct: Percent qualifying for CalWorks (income assistance).
   - meal_pct: Percent qualifying for free lunch.
   - computer: Number of computers.
   - testscr: Average test scores.
   - comp_stu: Number of computers per student.
   - expn_stu: Number of expenditure per student.
   - str: Student-teacher ratio.
   - avginc: Average income.
   - el_pct: Percent of English learners.
   - read_scr: Average reading score.
   - math_scr: Average math score.

   (a) Draw a scatterplot of test scores versus number of computers per student. Describe in words what you see.

   (b) Draw a scatterplot of test scores versus number of computers per student. Describe in words what you see.