1. Consider the following causal question: What is the effect of gender on earnings?

(a) What is the ideal experiment for answering this question? That is: i) what will you randomize; ii) what variable will you use to measure the effect of the randomization?

(b) How might you check if the randomization was done correctly?

(c) What are the potential outcomes for your experiment?

(d) What plausible causal channel(s) runs directly from the treatment to the outcome?

(e) 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. We will use a dataset from a randomised experiment conducted by Marianne Bertrand and Sendhil Mullainathan, who sent 4,870 fictious resumes out to employers in response to job adverts in Boston and Chicago in 2001.

The resumes differ in various attributes including the names of the applicants, and different resumes were randomly allocated to job openings. Some of the names are distinctly white sounding and some distinctly black sounding. The researchers collecting these data were interested to learn whether black sounding names obtain fewer callbacks for interviews than white names. (This introduction is taken from one of the problem set questions of Jorn-Steffen Pischke).

(a) Download and open the data set **bm.dta** from Github/Exercise.

(b) The variable “black” is a dummy variable that equals 1 when the resume has a black-sounding name. The key dependent variable is “call”, which equals 1 if the resume received a call back. Calculate the average callback rate for resumes with black-sounding names. What population parameter is approximated/estimated by this average?

(c) Calculate the average callback rate for resumes with white-sounding names. What population parameter is approximated/estimated by this average?

(d) Take the difference between the averages you calculated in the last two questions. What population parameter is approximated/estimated by this average?

(e) Regress “call” on “black”. How does the estimated coefficient on “black” compare with the difference you calculated in the last question?

(f) Regress “call” on “black” and “female”. How does the estimated coefficient on “black” compare with the coefficient you calculated in the last question?

(g) Regress “call” on “black” and every other control. How does the estimated coefficient on “black” compare with your last estimates? Using our ceteris paribus discussion from class, what can infer about the relationships between “black” and the other controls?

(h) What do you conclude from the results of the Bertrand and Mullainathan experiment?

3. This and the next question ask you to build a population from scratch.

a) Run the following commands. Interpret or guess what they do. Thus, beside each command you should have a one word explanation for what the command does.

```
ssc install bnormpdf

clear

set obs 500

matrix V = (1, .4 \ .4, 1)

matrix list V

drawnorm Y0 Y1, means(2,3) cov(V)

bnormpdf Y0 Y1, m1(2) m2(3) s1(1) s2(1) rho(.2) dens(Density)
```

b) Run the following commands. Interpret the picture you get (it may take a some minutes to render).

```
label variable Y0 "Potential Outcome without treatment"

 label variable Y1 "Potential Outcome with treatment"

graph twoway contourline Density Y0 Y1
```

(c) Generate a dummy variable that equals 1 with probability 0.5. Label this variable D. This is our hypothetical treatment assignment variable.

(d) Use Y0, Y1, and D to construct an observed outcome. Call this observed outcome Y.

(e) Generate a variable that equals Y1-Y0. Label this variable “Causal”.

(f) Summarize the variables “Causal” and Y0.

(g) Regress your Y on D. Compare your summary statistics for “Causal” and Y0 with your regression estimates. What do you learn?

4. Keep the population you have built. Now:

(a) generate a variable that equals D. Call it W.

(b) replace W with 1 if D is equal to 0 and Y is larger than 3.

(c) replace W with 0 if D is equal to 1 and Y is less than 2.

(d) regress Y (from the previous question) on D and W. Inspect the estimated coefficient on D, focusing on the comparison with the estimated coefficient on D in a regression that excludes W. What do you conclude and why?