Simulation Exercise: Potential Outcomes

Notes

- This exercise relies heavily on fake data simulation. Fake data simulation means that we generate random variables such that they resemble intuitive quantities that illustrate a point we want to make. Generation of random variables means taking draws from a given distribution, characterized by a (set of) parameter(s). For instance, the flip of a fair coin is a draw from a Bernoulli distribution with probability parameter $p = 0.5$. Rolling a fair dice is a draw from a categorical distribution with probability parameters $\mathbf{p}' = \left\{\frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{6}\right\}$. Randomly pointing anywhere on a 1m measurement tape is a draw from a uniform distribution with support $[0, 1]$. You get the idea.

- Consider the following fictitious setting: We have a sample of $N = 1000$ students. Our outcome of interest is knowledge about counterfactual causality, measured by students' test scores in a quiz on potential outcomes. The treatment of interest is taking an undergraduate class in research design.

  Students' potential outcomes under the control, $Y_i(0)$, are evenly distributed between 20 and 80. The individual treatment effects, τ_i, range between 0 and 20. Thus, the potential outcomes under the treatment, $Y_i(1)$, may range between 20 and 100. In this exercise, two forms of bias will be introduced if students self-select into the class. Their probability of selecting into the class is a direct function of their prior ability, i.e., of their potential outcomes under the control, $Y_i(0)$. Similarly, prior ability affects how much they learn (i.e., the size of their idiosyncratic treatment effects, τ_i) – students with a higher prior ability tend get more out of the class. Therefore, self-selection will result in both selection bias and differential treatment effect bias.

  In our first scenario, students are randomly assigned to treatment and control – pure chance determines whether they take an undergraduate class in research design or not. In the second, more realistic scenario, students self-select into the class. Here, we want to quantify the magnitude of selection bias and differential treatment effect bias under randomization and under self-selection.

Prompt:

1. Before you start, set a seed for reproducibility of your random variable generation (`set.seed()`).
2. Generate an integer, `N <- 1000L`, to be used in determining the length of the variables we create below.
3. Generate a variable of potential outcomes under the control, `Y_0`, containing random draws from a uniform distribution with support $[20, 80]$ (`runif()`).

4. Generate a variable of individual treatment effects, `tau`. As `tau` is a function of the potential outcomes under the control, we have differential treatment effects. We draw these from a normal distribution with some random error (`tau <- 0.2 * rnorm(Y_0, 2.5)`) and then truncate the distribution at 0 and 20.

5. Show a scatter plot with `Y_0` on the x-axis and `tau` on the y-axis. Briefly interpret the pattern you observe.

6. Generate a variable of potential outcomes under the treatment, `Y_1`, using the two variables you previously generated.

7. Generate a randomly assigned binary treatment indicator, `D`, which takes on values of either 0 or 1 from a Bernoulli distribution, which is a special case of the binomial distribution with size $= 1$, meaning we take one draw for each observation i (`rbinom`). Suppose everyone has equal probability of being assigned to either treatment or control group.

8. Show that the potential outcomes are independent of treatment status. Choose an appropriate test or statistic and give a brief interpretation.

9. Generate a variable `Y_obs` that takes the values of `Y_0` if `D==0` and the values of `Y_1` if `D==1`.

10. Run a regression of `Y_obs` on `D` to gauge the average treatment effect. Give a brief interpretation.

11. Next, generate a variable `D_sel` that indicates receipt of the treatment in a scenario of self selection. For this, generate a variable `prob_sel` equal to the potential outcomes under control divided by 100. This gives you each student's probability of selecting into the treatment. Use these probabilities to draw `D_sel` from a Bernoulli distribution.

12. Using the same test you chose above, test if the potential outcomes are independent of `D_sel` and give a brief interpretation.

13. Generate a variable `Y_obs_sel` that takes the values of `Y_0` if `D_sel==0` and the values of `Y_1` if `D_sel==1`.

14. Run a regression of `Y_obs_sel` on `D_sel` to gauge the average treatment effect under selection bias.

15. Specify the magnitude of both selection bias and differential treatment effect bias. Explain how the two differ conceptually and interpret the estimates.