Exercise: Understanding Potential Outcomes

1. Generate data on a population of 1,000 people. Specifically, create a variable (a vector) that randomly assigns these people to be male or female (50:50). *Hint: In R, try rbinom and in Stata, try rbinomial.*
2. Now we are going to simulate the potential outcomes - a measure of attitudes - *under control* (y_0) for our population. Create another variable of random normally-distributed values with mean of 5 and standard deviation of 1.
3. One problem with observational data is that potential outcomes are often correlated with other variables such as gender. Adjust your value of y_0 to add 1 (one) for all units who are male.
4. Now simulate potential outcomes *under treatment* (y_1) for all units. Define a *constant* treatment effect of $c = 2$ and create another vector $y_1 = y_0 + c$.
5. To compare our two sets of potential outcomes, plot two density charts on the same figure - one for y_0 and one for y_1.
6. Next, let us assume a specific **Treatment Assignment Mechanism** where men are more likely than women to receive treatment. First, create a temporary 'latent' variable which consists of two components added together: (i) $0.5 * x$ (i.e. half the value of the gender variable), and (ii) a random uniform value between zero and one. Finally, create a new vector D which is equal to one when the latent variable is larger than 0.75, and zero otherwise.
7. To show that gender and treatment - maybe we can think of treatment as low income and high income - are related, calculate the correlation between x and D.
8. What is the average of the *real* indvidual treatment effects based on the potential outcomes, $E(y_1 - y_0)$?
9. The Fundamental Problem of Causal Inference is that we *cannot* calculate (8.) above. Instead, we only observe one value: y_{obs}. Create a new variable y_{obs} which equals y_1 if $D = 1$ but which equals y_0 if $D = 0$.
10. Based on the observable data, run the basic regression of treatment (D) on observable outcomes (y_{obs}). Interpret the result. Is this an accurate estimate of the treatment effect that we assumed at the start?
11. Re-run all your code above but this time with $c = 0$ so we are assuming **NO** treatment effect. Run the regression in (10.) again - what is the result?
12. To see why, let's plot two density charts on the same figure - one for the distribution of observable y_{obs} for the treated group ($y_{obs}|D == 1$) and one for the distribution of observable y_{obs} for the control group ($y_{obs}|D == 0$).
13. Run your code again for $c = 0$, but this time assume a larger population of $N = 1,000,000$. Does that solve the problem?
14. For $c = 0$, run the regression of treatment on observable outcomes, but this time controlling for gender.