Assignment 7: Causal Inference

PSCI 8357, Spring 2016
March 24, 2016

This assignment must be turned in by the start of class on **Thursday, March 31**. You must follow the instructions for submitting an assignment.

Main Task

1. Generate $N = 500$ observations of the following data.
   - Four binary covariates, $X_1, \ldots, X_4$, independent of each other and distributed Bernoulli(0.5) (i.e., 50-50 chance of being 0 or 1).
   - A binary treatment variable, T, where

     $$\Pr(T = 1) = \frac{4 + X_1 + X_2 + X_3 + X_4}{12}.$$

   - A continuous response variable, Y, where

     $$Y = T + \exp(X_1 + X_2 + X_3 + X_4 + X_1 X_2 + X_3 X_4) + \epsilon$$

     with white noise error $\epsilon \sim N(0, 1)$.

   Under this model, there is a constant treatment effect of $\tau = 1$.

   Since you are simulating data at random, remember to use `set.seed()` at the start of your script so your results are reproducible.

2. Use a naïve difference of means test to estimate the average treatment effect. How does the estimate compare to the true ATE?

3. Use subclassification to estimate the average treatment effect. How does the estimate compare to the true ATE?

   In the unlikely but not-impossible event that there is a grouping with no variation in the treatment, just drop that grouping from the subclassification.

4. Use OLS of Y on (T, X_1, X_2, X_3, X_4) to estimate the average treatment effect. How does the estimate compare to the true ATE?

5. Use a modified form of subclassification to estimate the average treatment effect. Instead of grouping on the full combination of covariates, just group on the sum $X_1 + X_2 + X_3 + X_4$. How does the estimate compare to the true ATE?

6. Use a loop to repeat the first five steps $M = 1{,}000$ times. Calculate the (approximate) bias, standard error, and mean squared error of the three estimators. (Remember that $\text{MSE} = \text{Bias}^2 + \text{Std. Error}^2$.) Test the hypothesis that each estimator is unbiased. Which of the estimators is best overall?

7. What if, instead of being binary, each covariate X_j were uniformly distributed between 0 and 1? Based on the results of your analysis here, how would you choose to estimate the average treatment effect? (**Don't** run a new simulation—try to make an inference from the results with binary covariates.)

Visualization Challenge

Depict the *joint* relationship between each X variable and (1) the probability of treatment and (2) the expected value of the response.