Economics 696F, Causal Inference and Program Evaluation

Problem Set 1 Suggested Solutions

1. Matlab code and conclusions are given in the M-file `hw1_1.m`

2. The answer to this question uses the Matlab program, `ols_het.m`, which calculates OLS estimates with Eicker-White heteroskedasticity-robust standard errors.

   (a) Since treatment is randomly assigned, we use the simple differences in means estimator

   $$\begin{aligned}\hat{\tau} &= \bar{Y}_1 - \bar{Y}_0 \\ &= \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1-T_i) Y_i}{\sum_{i=1}^n (1-T_i)}.\end{aligned}$$

   This is numerically equivalent to the slope coefficient in the OLS estimator that regresses Y_i on a constant and T_i. To write this formally, let

   $$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad W := \begin{pmatrix} 1 & T_1 \\ \vdots & \vdots \\ 1 & T_n \end{pmatrix}.$$

   The OLS estimator is

   $$b := \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} := (W'W)^{-1} W'Y.$$

   Then $b_2 = \hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.
   The "usual" variance formula $s^2(W'W)^{-1}$ is not valid under heteroskedasticity. Instead, we can use the Eicker-White formula:

   $$V(\hat{b}) = n \cdot (W'W)^{-1} \left[\frac{1}{n} \sum_{i=1}^n W_i W_i' e_i^2 \right] (W'W)^{-1},$$

   where $W_i = (1, T_i)'$, and $e_i := Y_i - W_i'b$.
   Using `ols_het.m`, we get $\hat{\tau} = 1794.34$, with a heteroskedasticity-robust standard error of 669.32. The t-statistic is 2.68, which leads us to reject the null hypothesis of zero average treatment effect at the 5% level.

   (b) By construction, the treatment is independent of variables such as age (since it was randomly assigned). Therefore, we expect that if we treated age as the outcome, we would estimate close to a zero "treatment effect" of the program on age.
   The estimated effect is $\hat{\tau} = 0.76$, with a standard error of 0.68. Thus the t-test would not reject the null hypothesis of zero effect at conventional levels.

3. We continue to use `ols_het.m` to solve this problem.

   (a) The simple difference in means estimate is -15,204.78, which means that treated individuals earn $15,205 less, on average, than controls. Even though the standard error is fairly large at 656, this is very different from the experimental results. Clearly, the assumption that treatment is independent of potential outcomes is not valid for this data set. It appears that among the individuals who did not get the job training program, were many with relatively high earnings potential.

   (b) We assume unconfoundedness and a linear regression specification

   $$E[Y|T,X] = \beta_1 + T_i\beta_2 + X_i'\beta_3 + (T_i \cdot X_i)'\beta_4,$$

   We want to estimate the effect **on the treated**, which can be written as:

   $$\beta_2 + E[X|T=1]'\beta_4.$$

   So a natural estimator is

   $$\hat{\tau} = \hat{\beta}_2 + \left[\frac{\sum_{i=1}^n T_i X_i}{\sum_{i=1}^n T_i}\right]' \hat{\beta}_4.$$

   Note that we need to take the average of X among the treated, not for the whole sample.

   We get an estimate of $\hat{\tau} = 580.56$, which is closer to the experimental results than the simple difference in means.