PS Weeks 2 and 3, Econometrics 3

For week 2, you need to submit exercises 1 and 2 to the TA. For week 3, you need to submit exercises 3 and 4. You need to submit a document with your answers, and the dofile you used to perform the estimations/simulations. Each week, the TA will randomly pick two questions and grade them.

Exercise 1: measuring the effect of tracking on students' test scores

This exercise uses data from "Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. The American Economic Review, 101(5), 1739-1774." In this paper, the authors randomly assigned 108 primary schools with 2 1st grade classes in Kenya to a tracking group and to a non-tracking group. In the tracking group, the 2 1st grade classes in the school were formed based on students' ability: students faring below the school median in a test in the beginning of the year go to class A, while students faring above the school median go to class B. In the non-tracking group, classes A and B are randomly formed, and therefore both classes have students below / above the median.

The data set tracking.dta contains data on 5170 students in these 108 primary schools. schoolid is the unique identifier of their school. tracking is an indicator equal to 1 for students whose school was randomly assigned to the tracking group. bottomhalf is an indicator equal to 1 for students who fared below the median of their school in the beginning of first grade test. Finally, scoreendfirstgrade is students' standardized score in the end of first grade test.

Throughout the exercise, you need to bear in mind that the randomization took place at the school level, not at the student level. Therefore, your unit of observation in all of what follows should be schools, not students. Hint: the Stata command "collapse" might prove useful.

1. Estimate the effect of tracking on students' end of first grade test scores.

2. Use a 10% level randomization inference test to assess whether the finding of question 1 is robust or whether it rests on an unappropriate asymptotic approximation.

3. One might fear that tracking benefits strong students while harming weaker ones. Assess whether this is a legitimate concern.

Exercise 2: some properties of the ATE estimator in randomized experiments

1. Prove the following theorem:

Theorem 0.0.1 *Using draws from a uniform distribution to generate draws from other continuous distributions*

Let F denote a strictly increasing cdf. If U follows the uniform $[0,1]$ distribution, then the cdf of $F^{-1}(U)$ is F.

Hint: the cdf of $F^{-1}(U)$ if the function $x \mapsto G(x) = P(F^{-1}(U) \leq x)$. You need to show that $G(x) = F(x)$.

The following questions require using Stata.

2. Set the number of observations to 1000, create a variable $y(0) = U$, where U follows a $N(0,1)$ distribution, and create a variable $y(1) = 0.5y(0) + 0.5V + 0.2$, where V is a $N(0,1)$ random variable independent of U. $y(0)$ and $y(1)$ represent the potential outcomes of 1000 units that participate in a randomized experiment. Hint: it follows from the previous question that if ε follows a uniform distribution, $\Phi^{-1}(\varepsilon)$ follows a $N(0,1)$ distribution. The Stata command for Φ^{-1} is invnormal().

3. Given the way we created $y(0)$ and $y(1)$, is the effect of the treatment homogeneous or heterogeneous across units?

4. Compute ATE_{1000}, the average effect of the treatment in this fixed population of 1000 units. Store this number in a scalar.

5. Compute the correlation between $y(1)$ and $y(0)$ and the variance of $y(1) - y(0)$ in this finite population of 1000 units.

6. Write a 800 iterations loop, where in each iteration you:

   1. Create a variable containing a random number.
   2. Sort the 1000 observations according to that random number.
   3. Create a dummy variable D equal to 1 for the first 500 observations in the sorted data set.
   4. Create a variable $Y = (1 - D)y(0) + Dy(1)$.
   5. Regress Y on D, using the robust option. Check that the coefficient of D is equal to $\widehat{ATE}_{1000}$. Check that the variance of that coefficient is equal to $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$.
   6. Compute $IC(0.05)_{+} = \left[\widehat{ATE}_{1000} - 1.96\sqrt{\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}}, \widehat{ATE}_{1000} + 1.96\sqrt{\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}}\right]$, and the indicator $1\{ATE_{1000} \in IC(0.05)_{+}\}$.
   7. Store $\widehat{ATE}_{1000}$, $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$, and $1\{ATE_{1000} \in IC(0.05)_{+}\}$ in a matrix.

Compute the mean of $\widehat{ATE}_{1000}$ over those 800 replications, and compare it to ATE_{1000}. Compute the variance of $\widehat{ATE}_{1000}$ over those 800 replications, and compare it to the mean of $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$ over those replications. Compute the mean of $1\{ATE_{1000} \in IC(0.05)_{+}\}$. Explain your results in light of the lecture notes.

7. Write a 800 iterations loop, where in each iteration you:

1. Create a variable $y(0) = U$, where U follows a $N(0,1)$ distribution, and create a variable $y(1) = 0.5y(0) + 0.5V + 0.2$, where V is a $N(0,1)$ random variable independent of U, with 1000 observations each.

2. Create a variable containing a random number.

3. Sort the 1000 observations according to that random number.

4. Create a dummy variable D equal to 1 for the first 500 observations in the sorted data set.

5. Create a variable $Y = (1-D)y(0) + Dy(1)$.

6. Regress Y on D, using the robust option. The coefficient of D is equal to $\widehat{ATE}_{1000}$. The variance of that coefficient is equal to $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$.

7. Compute $IC(0.05)_{+} = \left[\widehat{ATE}_{1000} - 1.96\sqrt{\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}}, \widehat{ATE}_{1000} + 1.96\sqrt{\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}}\right]$, and the indicator $1\{0.2 \in IC(0.05)_{+}\}$.

8. Store $\widehat{ATE}_{1000}$, $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$, and $1\{0.2 \in IC(0.05)_{+}\}$ in a matrix.

Compute the mean of $\widehat{ATE}_{1000}$ over those 800 replications, and compare it to 0.2. Compute the variance of $\widehat{ATE}_{1000}$ over those 800 replications, and compare it to the mean of $\widehat{V}\left(\widehat{ATE}_{1000}\right)^{+}$ over those replications. Compute the mean of $1\{0.2 \in IC(0.05)_{+}\}$ (i.e. $1\{ATE_{1000} \in IC(0.05)_{+}\}$). Use the lecture notes to explain why the results change wrt to those you obtained in question 6.

Exercise 3: Some more properties of conditional expectations

Let X and Y be discrete random variables takink respectively n and m values denoted $x_1, ..., x_n$ and $y_1, ..., y_m$.

1. Show that
$$E(Y|X = x_k) = \frac{E(Y1\{X = x_k\})}{P(X = x_k)}$$

2. Show that if $Y \perp\!\!\!\perp X$, $E(Y|X) = E(Y)$.

Exercise 4: Estimating the average variance of independent but not identically distributed random variables (Midterm 2020)

Let $(Y_i)_{1 \leq i \leq n}$ be n independent but not identically distributed random variables. We seek to estimate $\frac{1}{n}\sum_{i=1}^{n} V(Y_i)$, the average of their variances. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the sample average of those variables. The goal of this exercise is to show that the sample variance of those variables is an upward biased estimator of their average variance.

1. Show that $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i)^2 - \frac{n}{n-1}(\overline{Y})^2$.

2. Show that $E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i)^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}V(Y_i) + \frac{1}{n}\sum_{i=1}^{n}(E(Y_i))^2\right)$.

3. Show that $E\left((\overline{Y})^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}V(Y_i) + \left(\frac{1}{n}\sum_{i=1}^{n}E(Y_i)\right)^2$.

4. Combine the results from the three questions above to show that

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right) = \frac{1}{n}\sum_{i=1}^{n}V(Y_i) + \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}\left(E(Y_i) - \frac{1}{n}\sum_{i=1}^{n}E(Y_i)\right)^2. \tag{0.0.1}$$

5. Use Equation (0.0.1) to finally prove that $E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right) \geq \frac{1}{n}\sum_{i=1}^{n}V(Y_i)$.

6. Find a necessary and sufficient condition to have that $E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right) = \frac{1}{n}\sum_{i=1}^{n}V(Y_i)$.