Princeton University **Spring 2021**
Department of Economics **ECO 518**

Exercise 10

Version 1 (April 16, 2021)

Mikkel Plagborg-Møller

Due on Canvas by 11.59 pm on Friday, April 30

The problem set should be submitted as a single PDF file on the course website. You may turn in one problem set per group of at most 3 students. You may discuss the exercises with any of your classmates. Please attach your code to the problem set. Use any software you like, although you learn the most by avoiding canned statistics/econometrics packages. If you find any errors or require clarification, please let me know right away.

Question 1

Consider the design of a completely randomized experiment without covariates. We plan to collect data $\{Y_i, D_i\}_{i=1}^N$ on outcome Y and randomly assigned binary treatment indicator D. Let $\bar{Y}_1$ and $\bar{Y}_0$ denote the sample averages of the outcomes in the treatment and control groups, respectively, and denote the corresponding sample sizes as N_1 and N_0. Assume that experimental subjects are drawn i.i.d. from some super-population. Denote the potential outcomes under treatment and control as Y_{1i} and Y_{0i}, respectively, and define $\mu_1 = E(Y_{1i})$, $\mu_0 = E(Y_{0i})$, $\sigma_1^2 = \text{Var}(Y_{1i})$, $\sigma_0^2 = \text{Var}(Y_{0i})$. For simplicity, assume that the normal approximation to the sampling distribution of $(\bar{Y}_1, \bar{Y}_0)$ holds exactly in finite samples:

$$\begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_0 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{N_1} & 0 \\ 0 & \frac{\sigma_0^2}{N_0} \end{pmatrix} \right).$$

To simplify notation, assume that the econometrician knows the true values of σ_1 and σ_0. Define $\alpha_{ATE} = \mu_1 - \mu_0$. Consider the usual two-sided t-test of $H_0 \colon \alpha_{ATE} = 0$ with significance level 5%.

i) Derive the power of the above-mentioned t-test at a specific alternative (μ_1, μ_0) for which $\mu_1 - \mu_0 = a \neq 0$.

ii) Show that the power function in (i) is strictly increasing in the quantity

$$\frac{|a|}{\sqrt{\sigma_1^2/N_1 + \sigma_0^2/N_0}}.$$

iii) Suppose we have available a fixed budget for the experiment, so the overall sample size $N = N_1 + N_0$ is fixed. Based on the result in (ii), how would you choose what fraction N_1/N of experimental subjects to randomly assign to the treatment group?

Question 2

We will use the data set `jtpa.csv` on the course website. Conducted in the late 1980s, the National JTPA Study is the largest randomized training evaluation ever undertaken in the United States. The study collected data on roughly 20,000 individuals in 16 different sites around the country. Eligible individuals who applied for JTPA services were first assigned to one of three different service groups (classroom training, on-the-job training, or a combination of the two). After this service group assignment, each participant had a 1/3 probability of being assigned to a control group. If assigned to the control group, the participant was not allowed to participate in JTPA training services. If assigned to the treatment group, the participant was allowed to participate in JTPA training services.

The data in `jtpa.csv` indicates whether the participant was treated (`treatment`=1 if treated) and what dollar amount the participant earned in total over the 30 months following random assignment (`earnings`). The data set includes adult participants only (youth participants have been dropped).

i) Test whether the difference in mean earnings between the treatment and control groups is significant. Allow for different variances by treatment status. What is the p-value for a two-sided test?

ii) Regress earnings on the treatment indicator, a constant, and age categorical dummies (already created in the dataset: age 22–25, age 26–29, age 30–35, age 36–44, age 45–54; the base category is age 55–78). Report the p-value of a two-sided test of the null of zero average treatment effect.

iii) Now pretend the researchers have not conducted the experiment yet, and they want to calculate how large the sample size needs to be in order to detect a $1000 increase in earnings over the following 30 months with a power of 0.80. Suppose they already have

an estimate of the standard deviation of earnings for the treated and for the control group from previous studies (use the sample standard deviations in the actual data set), and have decided to assign the participants to the treatment group with probability 2/3 and to the control group with probability 1/3. Using the results in Question 1, how large does the sample size need to be?

Question 3

We will exploit the research design in the article "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records" by Joshua D. Angrist (*American Economic Review* 1990). Angrist seeks to measure the average treatment effect of military service on later-in-life earnings. During the Vietnam war, the U.S. army used a lottery system for the draft. The 1970 lottery covered men born in 1944–50, the 1971 lottery covered men born in 1951, the 1972 lottery covered men born in 1952, and the 1973 lottery covered men born in 1953. Lottery numbers were randomly assigned. If a man drew a low lottery number, he was eligible to be drafted after further examination; men with high lottery numbers were not drafted. However, all men were allowed to volunteer for the army. 1972 was the last year in which men were actually drafted for the army, although lotteries and volunteering continued.

The data set `draft.csv` on the course website contains an extract of the data for the Angrist paper. The sample consists of 10,100 men born in the years 1950–1953, for which we observe four variables: `lwage` is log total wages for the years 1964–1984 (measured in 1000s of 1978 dollars); `yob` is the year of birth minus 1900; `draftelig` is a binary indicator that equals 1 if the individual's lottery number made him eligible to be drafted; and `veteran` is a binary indicator that equals 1 if the individual served in the military.

Consider an instrumental variables regression of log wages on a constant and veteran status, using draft eligibility as an instrument.

i) Report coefficients from the above-mentioned IV regression. Do the analysis separately by year of birth. Compute robust standard errors for all point estimates.

ii) To interpret the estimates from part (i) in the LATE framework, we require the exclusion restriction and monotonicity assumption to hold. What do these assumptions mean in the present application? Do the assumptions seem reasonable here? Be brief.

iii) What is a complier in the present application? Under the monotonicity assumption,

estimate the fractions of compliers, always-takers, and never-takers in each birth-year cohort.

iv) Do you expect the LATE in part (i) to be higher or lower than the average treatment effect for the entire population of young American men born in the early 1950s? Be brief.

Question 4

Consider an experiment with *one-sided non-compliance*. Y denotes an outcome of interest, $D \in \{0, 1\}$ denotes the indicator for actually receiving treatment, while $Z \in \{0, 1\}$ is an indicator for whether the subject was assigned to treatment by the researcher. Assume that it is only possible for an experimental subject to *receive* treatment when the researcher *assigns* that subject to treatment, although some subjects who are assigned to treatment end up choosing not to receive treatment. In other words, assume that $Z \geq D$ for everyone in the population.

Let Y_1 and Y_0 denote potential outcomes given actually *receiving* treatment or not receiving treatment, respectively. (We assume that these potential outcomes do not depend directly on treatment *assignment*.) Let D_1 and D_0 denote potential (received) treatment status given *assignment* to treatment or not being assigned to treatment, respectively. Assume that treatment *assignment* is randomized:

$$(Y_1, Y_0, D_1, D_0) \perp\!\!\!\perp Z.$$

Assume further that the observed outcome and treatment status are given by

$$Y = DY_1 + (1 - D)Y_0, \quad D = ZD_1 + (1 - Z)D_0.$$

Finally, assume $0 < \Pr(Z = 1) < 1$ and $\Pr(D_1 = 1) > 0$.

i) Consider the probability limit of the 2SLS estimator in a regression of Y on D and a constant, using Z as an IV. Show that this probability limit equals the average treatment effect for the treated (ATET). [Hint: Use the LATE theorem, but remember to show that it applies.]

Question 5

Consider the following hypothetical research design. There are two firms, firm A and firm B, located in the same town. At time $t = 0$, we randomly sampled 50 workers from each firm and surveyed their overall life satisfaction, measured on a scale from 1 (very unhappy) to 10 (very happy). Between time $t = 0$ and time $t = 1$, firm B went out of business and laid off all its workers. At time $t = 1$, we randomly sampled 50 workers from firm A and 50 former workers of firm B and surveyed their overall life satisfaction. The workers sampled at $t = 1$ are not necessarily the same as those sampled at $t = 0$. We are interested in estimating the causal effect of a layoff on individual life satisfaction.

i) Define a differences-in-differences point estimator of the average treatment effect on the treated. Be clear about what your notation means.

ii) Interpret the "parallel trends" assumption in the present context. Discuss a few reasons why the assumption may not hold. Be concrete and brief.

iii) What kinds of (realistically obtainable) data would help shed light on the plausibility of the "parallel trends" assumption? Be concrete and brief.