Advanced Econometrics

Problem Set 1

(Due Monday, October 21 in class)

1. In the potential outcomes framework, suppose that program eligibility is randomly assigned but participation cannot be enforced. To formally describe this situation, for each person i, z_i is the eligibility indicator and x_i is the participation indicator. Randomized eligibility means z_i is independent of (Y_{0i}, Y_{1i}) but x_i might not satisfy the independence assumption.

i. Explain why the difference in means estimator is generally no longer unbiased.

ii. In the context of a job training program, what kind of individual behavior(s) would cause bias?

解. i. 项目资格的随机发放实际上不构成 RCT，个体的参与行为实际影响 Y 的变化
令

$$x_i = \begin{cases} 1, & \text{参与了活动.} \\ 0, & \text{未参与活动.} \end{cases}$$

$$Y_i = \begin{cases} Y_{1i}, & x_i = 1 \\ Y_{0i}, & x_i = 0 \end{cases}$$

其条件期望为：

$$\begin{aligned} E(Y_i|x_i=1) - E(Y_i|x_i=0) &= E(Y_{1i}|x_i=1) - E(Y_{0i}|x_i=0) \\ &= E(Y_{1i} - Y_{0i}|x_i=1) + E(Y_{0i}|x_i=1) - E(Y_{0i}|x_i=0) \end{aligned}$$

已知此时 $E(Y_{0i}|x_i=1) - E(Y_{0i}|x_i=0)$ 为选择性偏误，代表参加项目和未参加项目人员本身的差距。由题目知 x_i 可能不满足独立性假设。
因此，$E(Y_{0i}|x_i=1) \neq E(Y_{0i}|x_i=0)$。
此时 $E(Y_i|x_i=1) - E(Y_i|x_i=0) \neq E(Y_{1i} - Y_{0i}|x_i=1)$。即均值估计量差异有偏。

ii. 由上文知选择性偏差的实际意义是个人特质由于选择造成的偏差。假设 Y 是个人的职业能力，由于在培训项目中参与无法被强制，那么具有较强职业能力的人员若偏向不参与，而初始能力较差的人员偏向于参与，那么就会导致 $E(Y_{0i}|x_i=1) - E(Y_{0i}|x_i=0)$ 值为负。从而缩小了 $E(Y_i|x_i=1) - E(Y_i|x_i=0)$ 的大小，使得估计量实际偏小。
证毕

2. The potential outcomes framework can be extended to more than two potential outcomes. In fact, we can think of the policy variable, w, as taking on many different values, and then $y(w)$ denotes the outcome for policy level w. For concreteness, suppose w is the dollar amount of a grant that can be used for purchasing books and electronics in college, $y(w)$ is a measure of college performance, such as grade point average. For example, $y(0)$ is the resulting GPA if the student receives no grant and $y(500)$ is the resulting GPA if the grant amount is \$500.

For a random draw i, we observe the grant level, $w_i \geq 0$ and $y_i = y(w_i)$. As in the binary program evaluation case, we observe the policy level, w_i, and then only the outcome associated with that level.

i. Suppose a linear relationship is assumed:

$$y(w) = \alpha + \beta w + \nu(0),$$

where $y(0) = \alpha + \nu$. Further, assume that for all i, w_i is independent of ν_i. Show that for each i, we can write

$$\begin{aligned} y_i &= \alpha_i + \beta w_i + \nu_i, \\ E(\nu_i|w_i) &= 0. \end{aligned}$$

ii. In the context of i, how would you estimate β (and α) given a random sample? Justify your answer.

iii. Now suppose w_i is possibly correlated with ν_i, but for a set of observed variables x_{ij},

$$E(\nu_i|w_i, x_{i1}, ..., x_{ik}) = E(\nu_i|x_{i1}, ..., x_{ik}) = \eta + \gamma_1 x_{i1} + ... + \gamma_k x_{ik}.$$

The first equality holds if w_i is independent of ν_i conditional on $(x_{i1}, ..., x_{ik})$ and the second equality assumes a linear relationship. Show that we can write

$$\begin{aligned} y_i &= \phi + \beta w_i + \gamma_1 x_{i1} + ... + \gamma_k x_{ik} + u_i, \\ E(u_i|w_i, x_{i1}, ..., x_{ik}) &= 0. \end{aligned}$$

What is the intercept ϕ?

iv. How would you estimate β (along with ϕ and the γ_j's in part iii? Explain.

解. i. 由题设知，解释变量和被解释变量之间满足线性关系，有 ols 估计知，有 $SSR(\alpha_i, \beta_i)$ 最小化。即

$$SSR(\alpha_i, \beta_i) = \sum_{i=1}^{N} (y_i - \alpha_i - \beta_i w_i)^2$$

对 SSR 取最小值，有一阶条件：

$$\frac{\partial SSR(\hat{\alpha}_i, \hat{\beta}_i)}{\partial \alpha_i} = 0, \frac{\partial SSR(\hat{\alpha}_i, \hat{\beta}_i)}{\partial \beta_i} = 0.$$

解得：

$$\begin{cases} \sum_{i=1}^{N} (y_i - \hat{\alpha} - \hat{\beta} w_i) = 0 \\ \sum_{i=1}^{N} x_i (y_i - \hat{\alpha} - \hat{\beta} w_i) = 0. \end{cases}$$

由第二个条件可以得出 $E(x_i(y_i - \hat{\alpha} - \hat{\beta} w_i)) = 0$。又因为 $\nu_i = y_i - \hat{\alpha} - \hat{\beta} w_i$。则有

$$E(\nu_i w_i) = 0.$$

又因为 ν_i 与 w_i 独立，所以有 $E(\nu_i w_i) = E(\nu_i)E(w_i) = 0$。可知 $E(\nu_i) = 0$。
得 $E(\nu_i|w_i) = E(\nu_i) = 0$

ii. 由上题一阶条件知：

$$\begin{cases} \sum_{i=1}^{N} (y_i - \hat{\alpha} - \hat{\beta} w_i) = 0 \\ \sum_{i=1}^{N} w_i (y_i - \hat{\alpha} - \hat{\beta} w_i) = 0. \end{cases}$$

令 $\bar{w} = \frac{1}{N}\sum_{i=1}^{N} w_i, \bar{y_i} = \frac{1}{N}\sum_{i=1}^{N} y_i$. 则有

$$\begin{cases} \bar{wy} = \hat{\alpha}\bar{w} + \hat{\beta}\bar{w^2} \\ \bar{y} = \hat{\alpha} + \hat{\beta}\bar{w} \end{cases}$$

求解上式得：

$$\begin{cases} \hat{\beta} = \frac{cov_N(w_i, y_i)}{Var_N(w_i)} \\ \hat{\alpha} = \bar{y} - \hat{\beta}\bar{w} \end{cases}$$

iii. 由上题知 $y_i = \alpha_i + \beta w_i + \nu_i$。令

$$\begin{cases} X = (x_{i1}, x_{i2}, x_{i3}, ..., x_{ik}) \\ \gamma = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3}, ..., \gamma_{ik}) \end{cases}$$

则 $E(\nu_i|w_i, x_{i1}, ..., x_{ik}) = \eta + \gamma_1 x_{i1} + ... + \gamma_k x_{ik} = \eta + X^T\gamma$。
根据条件期望的分解性质, 得 $\nu_i = E(\nu_i|w_i, x_{i1}, ..., x_{ik}) + \xi$ 且 $E(\xi|w_i, x_{i1}, ..., x_{ik}) = 0$
带入上式得出:

$$\begin{aligned} y_i &= \alpha_i + \beta w_i + E(\nu_i|w_i, x_{i1}, ..., x_{ik}) + \xi \\ &= \alpha_i + \beta w_i + \eta + X^T\gamma + \xi \end{aligned}$$

令 $\phi = \alpha_i + \eta, u_i = \xi$。则式子变为 $y_i = \phi + \beta w_i + X^T\gamma + u_i$.
且 $E(\xi|w_i, x_{i1}, ..., x_{ik}) = E(u_i|w_i, x_{i1}, ..., x_{ik}) = 0$

iv. 令 $X_i = (1, w_i, x_{i1}, ..., x_{ik})^T, \hat{\beta} = (\phi, \beta, \gamma_1, ..., \gamma_k)^T$。由 ols 推导最小化残差知，知 FOC 为

$$\sum_{i=1}^{k} X_i(y_i - X_i^T\hat{\beta}) = 0$$

解得：

$$\sum_{i=1}^{k} X_i y_i = (\sum_{i=1}^{k} X_i X_i^T)\hat{\beta}$$

则 $\hat{\beta} = \frac{\sum_{i=1}^{k} X_i y_i}{\sum_{i=1}^{k} X_i X_i^T)}$。其中 $\phi = \frac{cov(y_i, \tilde{w_i})}{Var(\tilde{w_i})}, \gamma_k = \frac{cov(y_i, \widetilde{x_{ik}})}{Var(\widetilde{x_{ik}})}$。$\tilde{w_i}, \widetilde{x_{ik}}$ 表示对其他回归元回归的残差。

证毕

3. The Current Population Survey (CPS) refers to any of the monthly surveys conducted by the US Census Bureau throughout the year, although the March CPS -considered the beginning of the annual survey cycle - is the most significant, and is the data used in this assignment. Broadly, the CPS collects cross-sectional employment data of the participating households, allowing for regression wherein the independent and dependent variables are associated with the same point in time. In this problem, we will explore the relationship between educational attainment on earnings. There are numerous sites you can download the CPS data from. One source among many is http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/

i. Create a figure with hourly wage plotted against educational attainment for men in the US between the ages of 30 and 40 in March of 2019.

ii. Estimate the CEF using OLS. Why is the CEF linear in this case and show that OLS will generate a consistent estimate of the CEF?

iii. We wish to estimate the causal effect on earnings of college attendance relative to only completing high school. Please use the framework of the Rubin Causal Model to assess if your comparison from question ii gives you a causal estimate, e.g., what is D_i, what is $E[Y_i(0)|D_i = 1]$, etc.

解. i. 在根据年龄，时间对于数据进行清洗过后得到 1717 符合条件的数据，同时去除存在缺失值的条目，得到 1275 条数据，绘制散点图得：

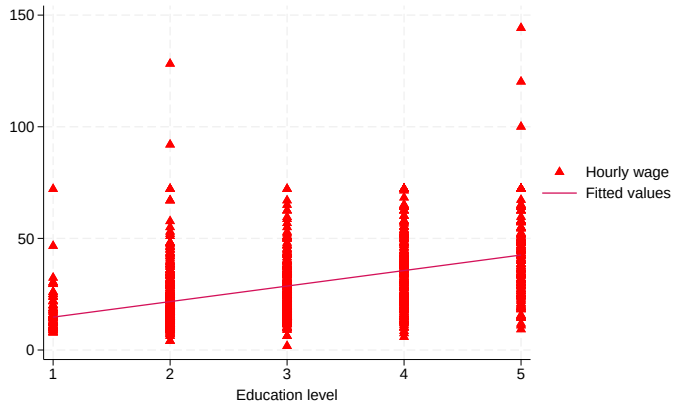

图 1: Scatter Plot

ii. 由 CEF 的分解公式知：$Y_i = E(Y_i|X_i) + \xi$。由于数据中每个回归元对应着对应着一个不同的参数，由此为饱和模型，则其 CEF 为线性的。根据线性条件期望函数定理，则总体回归函数也为线性的，条件期望的回归定理也为我们提供了 CEF 的最优线性近似。

$$\beta = \arg\min_b E(E[Y_i|X_i] - X_i b)$$

所以，由于 OLS 模型，对小时工资和教育水平进行回归，得：

4. We wish to evaluate the effect on annual earnings of the job training provided to participants in the National Supported Work (NSW) Demonstration study. In this study, a sample of men were randomly assigned to either receive or not receive job training. Baseline earnings were measured in 1975, which is the year before the they were randomly assigned to either the treatment or control group. Earnings were also measured in 1978, which is several years after the treatment started. More details on the NSW are available in Robert Lalonde, "Evaluating the Econometric Evaluations of Training Programs," American Economic Review, Vol. 76, pp. 604-620.

The dataset for this homework is called **nsw.dta**, which can be downloaded from https://users.nber.org/∼rdehejia/data/.nswdata2.html, and includes the following variables: treatment indicator ***treat*** (1 if treated, 0 if not treated), ***age*** (age in years), education (years of education), ***black*** (1 if black, 0 otherwise), ***hispanic*** (1 if Hispanic, 0 otherwise), ***married*** (1 if married, 0 otherwise), ***nodegree*** (1 if no high school degree, 0 otherwise), ***re75*** (earnings in 1975), and ***re78*** (earnings in 1978).

i. Calculate the ATE and the ATT of the policy usng a simple difference in means. How do the two values compare?

ii. Provide evidence that the randomization worked by comparing the means of the sample characteristics in the treatment and control group. Please create a clean table that includes columns with the means of each group, the difference between the two groups and the p-value of the difference. The table should be comprehensible on its own. Include a footnote for the table with a description of the dataset. A table of this form is usually the first one in an empirical study intended to recover a causal estimate. Is the table consistent with the randomization being correctly implemented?

iii. Create a table that presents your evaluation of the effect of the NSW experiment on earnings. In the first column present the raw difference in means between the treatment and control group then sequentially add covariates. In the last column include estimates from a regression with all covariates. Do the estimates change much as you add covariates? Why or why not? What does this tell you?

iv. Do you believe your estimates of the treatment effect are unbiased of the true treatment effect? Why or why not?

v. Estimate and plot the density functions of the 1978 earnings for the treatment and control group.

5. Duflo, Dupas, and Kremer (2011) investigate the impact of tracking (assigning students based on initial test score) on educational attainment in a randomized experiment. An extract of their data set is available on the Bruce Hansen textbook webpage in the file **DDK2011**.

In 2005, 140 primary schools in Kenya received funding to hire an extra first grade teacher to reduce class sizes. In half of the schools (selected randomly) students were assigned to classrooms based on an initial test score ("tracking"); in the remaining schools the students were randomly assigned to classrooms. For their analysis the authors restricted attention to the 121 schools which initially had a single first-grade class.

i. Do a regression of standardized test score (***totalscore*** normalized to have zero mean and variance 1) on tracking. Calculate standard errors using both the conventional robust formula, and clustering based on the school. Report and interpret the results.

ii. Do a regression of standardized test score on tracking, age, gender, being assigned to the contract teacher, and student's percentile in the initial distribution. (The sample size will be smaller as some observations have missing vari- ables.) Calculate standard errors using both the conventional robust formula, and clustering based on the school. Compare the two sets of standard errors. Which standard error changes the most by clustering? Which changes the least?

iii. How does the coefficient on tracking change by inclusion of the individual controls (in comparison to the results from i)?