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：

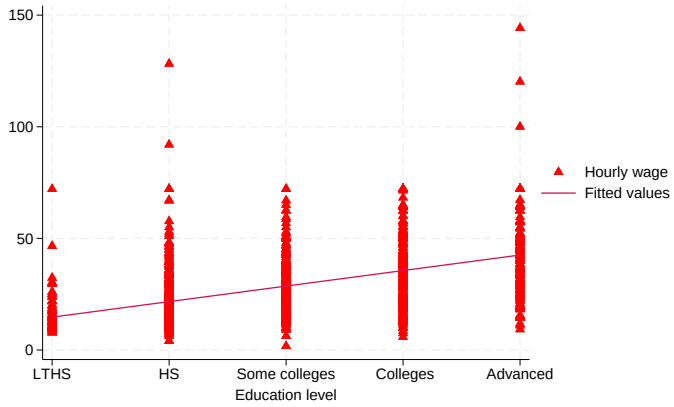

图 1: Scatter Plot

ii. 由 CEF 的分解公式知：$Y_i = E(Y_i|X_i) + \xi$。由于 CEF 的实际意义是在 X 给定的条件下，Y 的均值，因此对 X 即教育条件进行分组求均值，得出 CEF，进行 OLS 拟合如下图所示。可以看出 CEF 是线性近似的，那么由 CEF 的回归定理，最优线性估计以及线性条

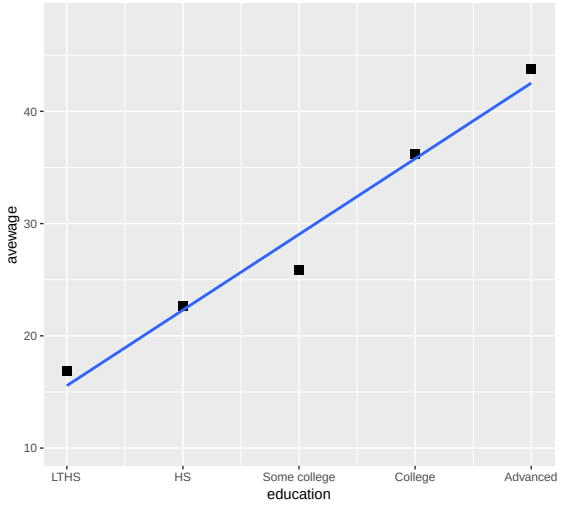

图 2: CEF Plot

件期望函数定理，总体回归方程也应该是线性近似的，且是对 Y 和 CEF 的最优线性估计量。

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

由上知，此种情况下 CEF 是线性的，则有 $Y_i = X_i\beta + e_i, \beta_{OLS} - \beta = (X'X)^{-1}X'e_i$。

通过变换：

$$\beta_{OLS} - \beta = (X'X)^{-1}X'e_i = (\frac{1}{N}X'X)^{-1}\frac{1}{N}X'e_i = (\frac{1}{N}\sum_{i=1}^{N}X_i'X_i)^{-1} - \frac{1}{N}\sum_{i=1}^{N}X_i'e_i$$

又因为在样本足够大时，样本均值依概率收敛于总体均值（总体矩）。则有 $\frac{1}{N}\sum_{i=1}^{N}X_i'X_i \xrightarrow{P} E(X_i'X_i)$, $\frac{1}{N}\sum_{i=1}^{N}X_i'e_i \xrightarrow{P} E(X_i'e_i)$ 又因为 $E(X_i'e_i) = 0$. 所以有：$\beta_{OLS} - \beta \xrightarrow{0}$。因此 OLS 将会产生 CEF 的一致估计。

iii. 根据条件期望的定义，有：

education	average wage
LTHS	16.82391
HS	22.62468
Some colleges	25.83010
College	36.16484
Advanced	43.75207

分别令教育水平依次从 1 到 5 进行排序，可知 D_i 表示所获教育的程度，有 ols 估计知容易知:

education	average wage
HS	22.62468
College	35.24900

$$D_i = \begin{cases} 1, & \text{上过大学} \\ 0, & \text{上高中未上大学} \end{cases}$$

且 $E(Y_i(0)|D_i = 1)$。表示上了大学的人如果未上大学的薪资。若无选择性偏误，那么其处理效应则是 12.6，如存在选择偏误，则无法测算。证毕

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.

解. i. 本题假设在三年期间不存在任何其他可能影响收入的事件发生，或者说假设培训对于收入在三年间的影响是最大的。对于 ATT 有 Rubin 模型知：ATE=$E(Y_{1i}|treat_i = 1)-E(Y_{1i}|treat_i = 0)$ 数据计算得：ATT= 2910.254 。而 ATE=2949.669 。因此 ATE>ATT。

ii. 通过数据的表格如下：

variables	means	difference	p-value
age0	24.45	0.18	0.72
age	24.63		
education0	10.19	0.19	0.14
education1	10.38		
black0	0.80	0.00	0.96
black1	0.80		
hispanic0	0.11	-0.02	0.42
hispanic1	0.09		
married0	0.16	0.01	0.70
married1	0.17		
nodegree0	0.81	-0.08	0.00
nodegree1	0.73		
re750	3026.68	39.42	0.92
re751	3066.10		

[1] P-value 小于 0.05 时证明在 5% 的置信区间内，差异显著

表 1: 随机化检验.

由上表知，除了学位，其他的特征的差异都不显著，P 值大于 0.05。因此可以认为随机化是有效的。

iii. 由上知，75 年的收入均值的差异不显著。根据 stata 做表得：

variables	treat==0	treat==1
mean	5090.048	5976.352
treat	886.3037	
age	882.2074	
education	831.039	
black	826.4462	
hispanic	824.3952	
married	820.4004	
re75	830.6734	
nodegree	806.5113	

表 2: 培训处理效果.

由表可以看出，表中的回归系数基本变化不大，这证明基本不存在选择偏误，表明明处理效应相对稳定，不受这些协变量的影响。实验的随机化很好。

iv. 由上两题知，实验的随机化水平很好，这意味着基本不存在选择偏误。那么实际的处理效应基本等于处理前后的收入的差距。由于初始的差异不显著，可以认为收入可以被很好的由是否培训的处理预测，即均值估计量差异基本无偏。

v. 依据数据，得出的 density functions 画图如下：

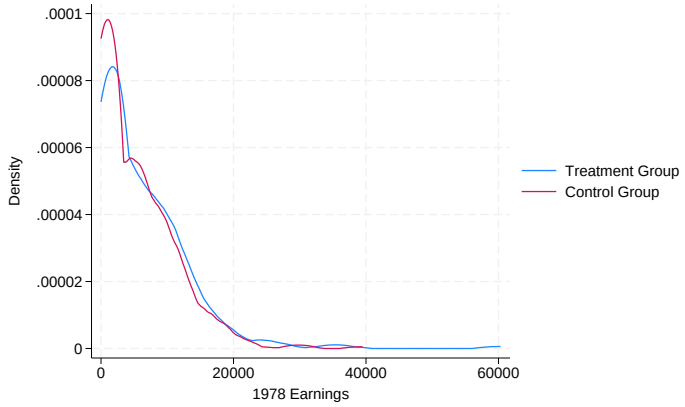

图 3: Density functions

证毕

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)?

解. i. 根据题意知，对于 totalscore 进行标准化。其次根据要求进行稳健回归，数据图如下所示：

standardize	coefficient	Robuststd.err	P
Tracking	0.1380913	0.0262102	0.00
cons	-0.0710354	0.0186418	0.00

表 3: 稳健标准误.

standardize	coefficient	Robuststd.err	P
Tracking	0.1380913	0.0772362	0.076
cons	-0.0710354	0.0543934	0.194

表 4: 聚类稳健的标准误.

由上表可以看出稳健的标准误以及聚类稳健的标准误情况下，系数基本没有差距，这表明分班对于测试成绩具有正向影响。同时也看出在学校层面的聚类标准误不显著，这可能是因为学校于分班存在某种相关性，导致聚类稳健的显著性下降。

ii. 由数据回归知，

standardize	coefficient	Robuststd.err	P
Tracking	0.1725117	0.0761819	0.00
agetest	-0.0408029	0.0133116	0.00
girl	0.0812035	0.0284988	0.00
etpteacher	0.17987574	0.0374764	0.00
percentile	0.0173172	0.0007203	0.00
cons	-0.729054	0.129734	0.00

表 5: 聚类标准误.

观察可知，聚类标准误在各个变量方向上基本都大于稳健标准误。在进行稳健标准误之后，Tracking 即是否有分班的影响。在聚类之后变化最大。

standardize	coefficient	Robuststd.err	P
Tracking	0.1725117	0.0240222	0.00
agetest	-00408029	0.0084928	0.00
girl	0.1798757	0.0240886	0.00
etpteacher	0.0173172	0.0237054	0.00
percentile	-0.0710354	0.0004246	0.00
cons	-0.729054	0.0809656	0.00

表 6: 稳健标准误.

iii. 由 i 题和 ii 题 comparison 知，在加入个体控制变量之后。Tracking 的系数有一个相对较大的上升，这表明了个体的控制变量会对于 Tracking 和得分之间 relationship 产生影响，存在选择偏误。使得处理效应估计的结果偏大。

证毕