Introductory Econometrics I – Spring 2024

Problem Set 3 – Due date: May 12

Last updated: April 26, 2024

Notes: Please submit a single PDF file containing your answers to all questions on Web-learning. For empirical questions, original codes and complete results need to be attached.

1. Consider the following regression:

$$y = \beta_0 + \beta_1 d + \beta_2 z + \beta_3\, d \cdot z + u,$$

where

- y is the personal income;
- d is a dummy (binary) variable for $female$ ($d = 1$ when the person is female, and $d = 0$ if the person is male);
- z is a dummy variable for $rural$ ($z = 1$ if the person lives in a rural area, and $z = 0$ if the person lives in an urban area).

We have a random sample $\{(y_i, d_i, z_i) : 1 \le i \le n\}$. The OLS regression estimators are denoted by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ (assuming the no perfect collinearity condition holds).

(a) Write the first-order conditions for the least squares regression problem.

(b) Define sample averages

$$\bar{y}_{11} = \frac{1}{n_{11}} \sum_{i=1}^{n} d_i z_i y_i, \qquad \bar{y}_{10} = \frac{1}{n_{10}} \sum_{i=1}^{n} d_i(1 - z_i) y_i,$$
$$\bar{y}_{01} = \frac{1}{n_{01}} \sum_{i=1}^{n} (1 - d_i) z_i y_i, \qquad \bar{y}_{00} = \frac{1}{n_{00}} \sum_{i=1}^{n} (1 - d_i)(1 - z_i) y_i$$

where n_{kl} denotes the number of persons with both $d_i = k$ and $z_i = l$, for $k, l \in \{0, 1\}$. How do you interpret the sample averages defined above?

(c) Show that

$$\sum_{i=1}^{n} d_i z_i (y_i - b_0 - b_1 d_i - b_2 z_i - b_3 \cdot d_i \cdot z_i) = 0$$

where

$$b_0 = \bar{y}_{00}, \quad b_1 = \bar{y}_{10} - \bar{y}_{00}, \quad b_2 = \bar{y}_{01} - \bar{y}_{00}, \quad b_3 = (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00}).$$

[Hint: use the fact that $d_i^2 = d_i$, $z_i^2 = z_i$ and $n_{11} = \sum_{i=1}^{n} d_i z_i$.]

(d) Now, *assume* the zero conditional mean condition: $\mathbb{E}[u|d, z] = 0$. In part (c), you actually show b_0, b_1, b_2, and b_3 satisfy one of the first-order conditions for OLS. In fact, it can be shown that the other first-order conditions are also satisfied. That means $\hat{\beta}_j = b_j$ for $j = 0, 1, 2, 3$. Use this fact to show

$$\beta_0 = \mathbb{E}[y|d = 0, z = 0],$$
$$\beta_1 = \mathbb{E}[y|d = 1, z = 0] - \mathbb{E}[y|d = 0, z = 0],$$
$$\beta_2 = \mathbb{E}[y|d = 0, z = 1] - \mathbb{E}[y|d = 0, z = 0],$$
$$\beta_3 = (\mathbb{E}[y|d = 1, z = 1] - \mathbb{E}[y|d = 1, z = 0]) - (\mathbb{E}[y|d = 0, z = 1] - \mathbb{E}[y|d = 0, z = 0]).$$

[Hint: Under the imposed conditions, we know $\hat{\beta}_j$ is unbiased for β_j for $j = 0, 1, 2, 3$. Take expectation of b_j conditional on $\{(d_i, z_i) : 1 \leq i \leq n\}$.]

(e) Use your answer to part (d) to explain the statistical meaning of the OLS estimator $\hat{\beta}_3$ (what does it really estimate?).

(f) Describe how to test the null hypothesis that the (population) average income of rural females does not differ from that of rural males at the 5% significance level.

(g) Describe how to test the null hypothesis that the (population) average income of females does not differ from that of males in both rural and urban areas at the 5% significance level.

2. (**Regression on a binary variable, revisit**) Consider again the simple regression on a binary variable (you studied it in Problem Set 1):

$$y = \beta_0 + \beta_1 d + u,$$

where y is the outcome of interest, and $d \in \{0, 1\}$ is a dummy (binary) variable indicating a "treatment". If a person is treated, then $d = 1$, and if not, $d = 0$. There is a random sample $\{(y_i, d_i) : 1 \leq i \leq n\}$. Let $n_1 = \sum_{i=1}^n d_i$ and $n_0 = n - n_1$. You have shown that the least squares estimator $\hat{\beta}_1 = \frac{1}{n_1} \sum_{i=1}^n d_i y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - d_i) y_i$, i.e., the difference in means between the treated and untreated groups.

Now assume everyone could have a *potential* outcome $y(1)$ if she had been treated and a potential outcome $y(0)$ if not treated. Then, the observed outcome y can be written as $y = dy(1) + (1 - d)y(0)$, i.e., we observe $y(1)$ if a person is treated and $y(0)$ otherwise (but we cannot observe both). We are interested in the population average treatment effect $\tau_{\mathtt{ATE}} := \mathbb{E}[y(1) - y(0)]$.

(a) Let $p_1 = \mathbb{P}(d = 1)$. Show that

$$\mathbb{E}[\hat{\beta}_1] - \tau_{\mathtt{ATE}} = \Big(\mathbb{E}[y(1)|d = 1] - \mathbb{E}[y(1)|d = 0]\Big)(1 - p_1) + \Big(\mathbb{E}[y(0)|d = 1] - \mathbb{E}[y(0)|d = 0]\Big)p_1.$$

[Hint: use law of iterated expectation and your answers to Q1 of Problem Set 1.]

(b) Do you think $\hat{\beta}_1$ is unbiased for $\tau_{\mathtt{ATE}}$? What if we also assume $\mathbb{E}[u|d] = 0$? (Does this change your answer?) Explain why.

(c) In this context, we are actually interested in the following regression:

$$y = \beta_0' + \tau_{\mathtt{ATE}} \cdot d + u',$$

where the slope on d is exactly the parameter we want to identify. Use the definitions of y, $y(1)$ and $y(0)$ to give explicit expressions of β_0' and u'. [Hint: decompose $y(1)$ and $y(0)$ into the (non-random) population mean and an (random) error term.]

3. (**Including Control Variables**) Suppose we want to estimate the causal effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized test score (say, *gaokaoScore*) and high school GPA (*hsGPA*) are also available.

(a) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$.)

(b) Should *gaokaoScore* and *hsGPA* be included as explanatory variables? Explain.

4. (**Data exercise**) Policy makers are interested in examining factors affecting the smoking behavior. They collect a data set about individual smoking behavior, including the following variables:

- `id`: individual index
- `age`: age of an individual
- `agesq`: age square
- `cigs`: number of cigarettes smoked per day
- `restaurn`: whether the individual lived in a city which requires no smoking in restaurants (0=no, 1=yes)
- `educ`: years of education

Please answer the following questions using the dataset `smoking.dta`:

(a) Create a new variable indicating age group named `agegrp`, which takes the following value:

$$agegrp = \begin{cases} 0, & \text{if age } \leq 30 \\ 1, & \text{if age } \in (30, 50] \\ 2, & \text{if age } \in (50, 70] \\ 3, & \text{if age } > 70. \end{cases}$$

Calculate the average of `cigs` for each age group. Do you think `age` and `cigs` has a monotonic relationship? [Hint: use the Stata command `tabstat cigs, by(agegrp) stat(mean)`.]

(b) Estimate the following regression by OLS:

$$cigs = \beta_0 + \beta_1 agegrp_1 + \beta_2 agegrp_2 + \beta_3 agegrp_3 + \beta_4 restaurn + u,$$

where for each $j = 1, 2, 3$, $agegrp_j$ is a dummy variable that equals 1 if $agegrp = j$ and 0 otherwise. Explain your estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$.

Now, estimate the following regression model instead using OLS:

$$cigs = \theta_0 + \theta_1 age + \theta_2 age^2 + \theta_3 restaurn + v,$$

(c) What is the marginal effect of age on cig? According to regression results, at what point does the marginal effect of age on $cigs$ change from positive to negative? (Round your answer to the nearest integer.)

(d) Explain the meaning of θ_3.

(e) Policy makers are interested in examining whether the partial effect of education on smoking is different for individuals living in cities with no-smoking mandate. Estimate the following regression model using OLS:

$$cigs = \gamma_0 + \gamma_1 educ + \gamma_2 restaurn + \gamma_3 restaurn \cdot educ + e.$$

Write out the expression for $\frac{\partial E(cigs)}{\partial educ}$ when $restaurn = 0$ and $restaurn = 1$. How do you understand the meaning of γ_3?

(f) Is γ_3 significant at 5% level?