Problem Set 2 (ver. 2024)

Econometrics 1

Fall semester 2024

Many thanks to Hoshino-sensei

Exercise 1 Show the following equalities hold: (i) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \overline{X}_n)$; and (ii) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \overline{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i(X_i - \overline{X}_n)$.

Exercise 2 Consider the following simple regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, \quad i = 1, \ldots, n.$$

Derive the OLS estimator for (β_0, β_1).

Exercise 3 Consider a regression model that has no intercept term:

$$Y_i = X_i\beta_1 + \epsilon_i, \quad i = 1, \ldots, n.$$

Derive the least squares estimator for β_1.

Exercise 4 Consider the following simple regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, \quad i = 1, \ldots, n.$$

Here, the explanatory variable X is normalized so that $\overline{X}_n = 0$.[1] Derive the OLS estimator for (β_0, β_1).

Exercise 5 Let $(\widehat{\beta}_{0n}, \widehat{\beta}_{1n})$ be the OLS estimator of

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, \quad i = 1, \ldots, n.$$

The prediction error (i.e., residual) for each i is $\widehat{\epsilon}_i = Y_i - \widehat{\beta}_{0n} - X_i\widehat{\beta}_{1n}$. Show that (i) $\sum_{i=1}^{n}\widehat{\epsilon}_i = 0$ and (ii) $\sum_{i=1}^{n} X_i\widehat{\epsilon}_i = 0$ hold.

Exercise 6 Consider the following simple regression model:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i, \quad i = 1, \ldots, n$$

where X_i is a dummy variable that takes either 0 or 1. Let $\overline{Y}_d$ be the average of Y over the observations with $X_i = d$: i.e., $\overline{Y}_d = n_d^{-1}\sum_{i=1}^{n}\mathbf{1}\{X_i = d\}Y_i$, where $n_d = \sum_{i=1}^{n}\mathbf{1}\{X_i = d\}$. Show that the OLS estimator for β_1 is obtained by $\overline{Y}_1 - \overline{Y}_0$.

[1] This is always possible by re-defining X_i by X_i <- $X_i - \overline{X}_n$.

Exercise 7 Let X_i be a dummy variable that takes either 0 or 1. In the following regression model, β_1 and β_2 are not separately estimable.

$$Y_i = \beta_0 + X_i\beta_1 + \exp(X_i^2)\beta_2 + \epsilon_i, \quad i = 1, \ldots, n.$$

Explain why.

Exercise 8 Let X_2 be determined by $X_2 = a + bX_1$, where a and b are some constants. In the following regression model, β_1 and β_2 are not separately estimable.

$$Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i, \quad i = 1, \ldots, n.$$

Explain why.

Exercise 9 Suppose you have data $\{(X_i, Y_i) : i = 1, \ldots, 500\}$, and the scatter plot of the data is given in "reg.png". What regression model would you employ to best explain the relationship between X and Y?

Exercise 10 Suppose you would like to test the hypothesis that *the value of acquiring computer skills is greater (in terms of relative wages) in developing countries than in developed countries.* If you have data on *wage*, *programming experience* (years), and other control variables for workers in a developed country A and those in a developing country B, how would you examine your hypothesis?

Exercise 11 Suppose that the distance between the star Q and the Earth is α light years. You have data $\{Y_1, Y_2, \ldots, Y_n\}$ of n independent trials of measuring the distance with the same equipment. Each trial entails some measurement error, but we can assume that the errors have zero mean.

1. Derive the least squares estimator for α, say $\widehat{\alpha}_n$.
2. Show that $\widehat{\alpha}_n$ is an unbiased estimator of α.
3. Show that $\widehat{\alpha}_n$ is a consistent estimator of α. (You can describe either mathematically or in words.)

Exercise 12 The following is the regression result on a sample of 1000 apartment houses in a Japanese city (standard errors in parentheses):

$$\widehat{rprice} = \underset{(9.9)}{71.3} + \underset{(0.3)}{1.4} \cdot area - \underset{(0.5)}{1.9} \cdot age + \underset{(0.4)}{0.9} \cdot renov$$

where *rprice* is the monthly rental price (1,000JPY), *area* is the area size (m^2) of the house, *age* is the age of the apartment (years), and *renov* is a dummy variable *renov* = 1 if the apartment underwent a renovation.

1. Suppose that an apartment owner renovates the apartment. What is the predicted increase/decrease in the rental price of a house in the apartment?

2. Suppose that, instead of measuring *rprice* and *area* in 1,000JPY and m^2, these variables are measured in 100JPY and $10m^2$. What is the coefficient estimate of *area* from this new regression?

3. Is the hypothesis *renovated and un-renovated apartments have the same economic value* rejected at the 5% significance level? If yes (no), what about the significance level at 1% (10%)?

Exercise 13 The following table shows the linear regression result on a sample of 500 students in a university. The dependent variable is the students' GPA score.

Variable	Coefficient	Std. Error
Male	-0.21	0.30
MothEduc	0.15	0.04
Commute	0.32	0.27
BF GPA	0.59	0.21
(Intercept)	2.10	0.57

Here, *Male* is a dummy variable *Male* = 1 for male students, *MothEduc* is the student's mother's education level in years, *Commute* is the commuting time to the university measured in hours, and *BF GPA* is the student's best friend's GPA score.

1. Can you conclude that female students outperform male students? Why?

2. Can you conclude that commuting time does not have any effect on one's GPA? Why?

3. Suppose that, instead of measuring *Commute* in hours, it is now measured in minutes. What is the coefficient estimate of *Commute* from this new regression?

4. The variable *BF GPA* is clearly an endogenous variable. (i) Why? (ii) Give an example of an instrumental variable (IV) for *BF GPA*, and explain why it is a valid IV.

Exercise 14 Explain (mathematically) why the OLS estimator is biased in the presence of endogenous regressors.

Exercise 15 Give specific examples of endogeneity bias caused by (i) omitted variables and (ii) simultaneity, separately for each (other than those mentioned during the lecture).

Exercise 16 Suppose that we would like to estimate the causal impact of aircraft noise on land prices by estimating a regression model: $\textit{land price} = \beta_0 + \textit{noise level}\beta_1 + \epsilon$. In this regression model, the aircraft noise variable is likely endogenous. Explain why, and give an example of a valid instrumental variable in this analysis.

Exercise 17 Suppose that we would like to estimate the causal impact of regional unemployment on crime rate by estimating a regression model: $\textit{crime rate} = \beta_0 + \textit{unemployment rate}\beta_1 + \epsilon$. In this regression model, the unemployment rate variable is likely endogenous. Explain why, and give an example of a valid instrumental variable in this analysis.

Exercise 18 In the field of international economics, there is a model called the *gravity model*, which assumes that the volume of economic trade between countries i and j, say Y_{ij}, is determined through the following function:

$$Y_{ij} = A\frac{\mathrm{GDP}_i^{\beta_1}\mathrm{GDP}_j^{\beta_2}}{\mathrm{dist}_{ij}^{\beta_3}}u_{ij},$$

where A is a constant term, dist_{ij} is the geographical distance between i and j, and u_{ij} is an error term with expectation one. Explain how you would estimate the parameters β_1, β_2, and β_3.