Problem Set 1

Yuta Toyama

Last updated: 2024-04-09

Due date: April 28 (Sun), 11pm (Japanese Standard Time)

Remarks

- You are allowed (and encouraged) to form a study group consisting of up to three members.
- Submit your report on Moodle. If you collaborate in a group, one submission per group will suffice. Kindly include the names of all group members.
- You should submit one single PDF file that includes both your report (as the main body) and programming code printout (as an appendix).
- Your report should be self-contained, containing all necessary information for me to understand your answer without consulting external sources, except for the problem instructions and lecture notes.
- Avoid including programming code in the main body. Instead, utilize words and equations to explain what you do.
- Present any results in a table format. Avoid using statistical program printouts that lack formatting.
- Include your programming code printout in an appendix, separate from the main body. Ensure that you include comments explaining your methodology in the programming code.
- [Important] Familiarize yourself with the rules surrounding plagiarism. Any form of plagiarism will not be tolerated, including copying solutions from online sources, previous year's solutions, classmates, and so on.

Question 1: Linear Regression

Use `MROZ_mini.csv` file for this question.

1. Consider the following linear regression model.

$$\log(wage)_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

   where $E[\epsilon_i | educ_i] = 0$ for now. Estimate the coefficients by OLS using the matrix algebra for linear regression. **Do not use the pre-existing package or command.**

2. Estimate the coefficient by numerically minimizing the objective function of the OLS estimator.

Specifically, you need to minimize the following objective function

$$J(\theta) = \sum_{i=1}^{N} (\log(wage)_i - \beta_0 - \beta_1 educ)^2$$

where $\theta = (\beta_0, \beta_1)$. Verify that your numerical solution is the same as (or sufficiently close to) the analytical solution you obtained above.

3. Calculate the estimates of the asymptotic standard errors of the coefficients using the matrix algebra. You can refer, for example, Proposition 2.1 of Hayashi (2000, pp.113). Do not report the number from the pre-existing package or command, though it would be useful to compare your answer with them so that you get the right number.

4. Do you think the mean independence assumption $E[\epsilon_i | educ_i] = 0$ is likely to hold? If not, what is the bias of the OLS estimator?

5. Consider the instrumental variable regression. The variable $fathereduc_i$ is the years of schooling of his or her father. Discuss the validity of $fathereduc_i$ as an instrumental variable for $educ_i$.

6. Run the IV regression using $fathereduc_i$ as an instrument for $educ_i$. Again, you should not use the pre-exising package for the instrumental variable regression. Chapter 5 of Wooldridge (2010) might be a good reference, though you can refer other econometrics textbook such as Hayashi, Hansen, etc. Prepare a table for regression results including both IV and OLS estimates. The table should include standard items such as coefficients, (asymptotic) standard errors, the sample size, etc.

Question 2: Discrete Choice Model

Use `data_KinokoTakenoko.csv` for this exercise.

I conducted a survey about chocolate snack (*Kinoko no yama* and *Takenoko no sato*) in my undergraduate IO course in Spring 2020. Please refer the lecture slide for the details.

The variables in the dataset are defined as follows.

Variable name	Description
id	ID for respondent. Each respondent (student) answers five questions.
occasion	Choice occasion (ignore "X" at the beginning).
choice	1: Kinoko, 2: Takenoko, 0: Other (outside option) |

The table reports summary statistics of the survey.

	Other	Kinoko	Takenoko
Q1: (Kinoko, Takenoko) = (200, 200)	0.41	0.21	0.39
Q2: (Kinoko, Takenoko) = (170, 200)	0.27	0.53	0.19
Q3: (Kinoko, Takenoko) = (240, 200)	0.37	0.08	0.55

	Other	Kinoko	Takenoko
Q4: (Kinoko, Takenoko) = (200, 250)	0.50	0.38	0.12
Q5: (Kinoko, Takenoko) = (200, 180)	0.24	0.11	0.65

Consider the following discrete choice model.

- Choice occasion $k = 1, \cdots, 5$
- Consumer i choose alternative j that gives the highest utility

$$
\begin{aligned}
U_{i,k,Kinoko} &= \alpha_{i,Kinoko} + \beta_i (y - p_{Kinoko,k}) + \epsilon_{i,k,Kinoko} \\
U_{i,k,Takenoko} &= \alpha_{i,Takenoko} + \beta_i (y - p_{Takenoko,k}) + \epsilon_{i,k,Takenoko} \\
U_{i,k,outside} &= \beta_i y + \epsilon_{i,k,other}
\end{aligned}
$$

  - $p_{j,k}$: price, y: income (300 JPY)
  - $\epsilon_{i,j,k}$: idiosyncratic shock, following type 1 extreme value distribution.
  - $\theta_i \equiv (\alpha_{i,Kinoko}, \alpha_{i,Kinoko}, \beta_i)$: preference parameters
    - $\alpha_{i,Kinoko}, \alpha_{i,Kinoko}$ follows the normal distribution.
    - β_i follows log-normal distribution.

Suppose for now that θ_i is a fixed parameter. That is we estimate $\theta = (\alpha_{Kinoko}, \alpha_{Takenoko}, \beta)$.

1. Suppose hypothetically that there is no outside option in the choice set so that respondants only answer to either buying Kinoko or Takenoko. Can you estimate all parameters? Explain.
2. Although the dataset does not have income information y_i, we can still estimate the model. Explain why.
3. Suppose that we only have one choice occasion, instead of five occasions. Can we estimate all parameters? Explain why.
4. Estimate logit model by maximizing likelihood function. Note that individual likelhood function is

$$L_i(\theta | \{y_{i,k}\}_{k=1}^{5}) = \prod_{k=1}^{5} P_{i,k} \left(j = y_{i,k} | \theta \right)$$

and the likelihood function is

$$L(\theta) = \prod_{i=1}^{N} L_i(\theta).$$

I recommend you to maximize the log-likelihood (rather than the original likelihood). **Do not use preexisting packages.**

Advice on Numerical Optimization

You need to use numerical optimization in Question 2. To do this, you first define the objective function that you want to maximize (or minimize). Then you apply numeral optimization tool (such as `optim` in R and `fminunc` in Matlab) to the function you defined.

You also need to choose the initial value in optimization. Choose it wisely. For example, you can run a linear regression by treating the choice variable as the dependent variable (i.e., regression “Kinoko” dummy on price) and use the estimated coefficient as an initial value. Also, you should check whether your optimization result is sensitive to your initial values.

Question 3: Monte Carlo simulation

We often need to use Monte Carlo simulation to approximate the integral. One such example is to calculate the choice probability in the random coefficient logit model. In this question, we conduct a simple exercise of Monte Carlo simulation.

Consider the random variable $X \sim N(\mu, \sigma^2)$ and you want to calculate $E[X^2]$.

Throughout the question, set $(\mu, \sigma^2) = (2, 2)$

1. Calculate the value of $E[X^2]$. You can do this analytically.
2. We now calculate this value by a Monte Carlo simulation. We approximate the integral by

$$E[X^2] = \int x^2 f(x) df \approx \frac{1}{R} \sum_{r=1}^{R} (x_r)^2$$

   where x_r is the random draw generated by the distribution $N(\mu, \sigma^2)$. R is the number of random draws, which you can choose. Report the value and discuss how it is close to the analytical value. How does it depend on the choice of R?

Advice

- Most scientific programming software has a random number generator (such as `rnorm` in R).
- I recommend you to draw the random numbers from the standard normal distribution $N(0, 1)$ and transform the draws by $\mu + \sigma \nu_r$ where ν_r is drawn from the standard normal.
- Set the seed of random number generators. Otherwise, you will get a different result every time you run a code.
- If you wish, you can use other techniques (such as importance sampling, Halton draw, quadrature, etc).