Introduction to Metrics: Homework 2

Business School, Nanjing University

Haocheng Hu & Zhaopeng Qu

4/24/2025

目录

1 Learning Objectives

- Review some key concepts from the lectures
- Learn how to use R to run discrete choice models,sample selection models,matching,etc.

2 Due Date and Submission Guidelines

- Due by May 13, 2:00am.
  - Late submissions will incur a penalty of "10% per hour"
- Upload your report (**including PDF or docx files**) along with all source files (such as `.qmd`, `.rmd`, and `.R` files for **R** users) to the **教学立方** system. For guidance on using this platform, please refer to the **Student User Guide**
- "All files" should include your formal report in *Word* or *PDF* format, *Rmarkdown* (`.Rmd`) or *Quarto* (`.qmd`) files used to generate the report, as well as R scripts (`.R` files for **R** users) used to clean the survey data for the final question.
- **Language**
  - *English or Chinese* (Writing in English will earn a small bonus)
- **Important Rule**: Plagiarism Will NOT Be Tolerated! If our TA finds sufficient evidence that your homework is **very similar** to another student's work or that you have used AI tools **inappropriately**, you will receive a score of **ZERO**!

3 Theoretical Exercises(40 points)

1. Suppose our multiple OLS model is:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_k X_{i,k} + u_i, i = 1, ..., n$$

(1)Under 1-4 OLS assumptions plus **homoskedasticity** assumption, please derive the variance of the OLS estimator of β_1, thus the expression of $\text{Var}(\widehat{\beta}_1)$(Hint: you can use the Frisch-Waugh-Lovell theorem to derive the expression)

(2) Discuss if following factors *decrease*, then how the $\sigma^2_{\hat{\beta}_1}$ of $\widehat{\beta}_1$ change and explain the reason.

- The sample size n **decreases**
- The variance of the error term u_i **decreases**
- The sample variance of $X_{i,1}$ **decreases**
- The R-square of the X_1 on other independent variables **decreases**

2. Suppose our simple OLS model is:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + u_i, i = 1, ..., n$$

Under 1-3 OLS assumptions plus **homoskedasticity** assumption and normality of u_i, please derive the distribution of $\widehat{\beta}_1$, please prove MLE estimator and OLS estimator are equivalent in this case.

3. Please read the content on the slides about Jia,Lan and Miguel(2021) carefully[or the paper] and answer the questions below:

(1) Look at **Table 3A** and only focus on column (2) and column (3),what is the key information that you can get from the results? Please explain in detail how to explain the estimated coefficients for `Cadre Parent` and `Entrepreneur Parent`? And evaluate whether those coefficients are **statistically and economically** significant or not?

(2) Look at **Table 3B** and only focus on **column (3)** in Panel A `Work in government`, what is the key information that you can get from the results? Please explain in detail how to explain the estimated coefficients for `Cadre Parent` and `Entrepreneur Parent`? And evaluate whether those coefficients are **statistically and economically** significant or not?

(3) Look at equation(2) in the paper, which parameter is the coefficient of interest? Please explain in detail how to interpret the coefficient of interest?

(4) Look at Equation(2) and Table4, focus on column (3) and column (6), what is the key information that you can get from the results? And evaluate whether those coefficients are **statistically and economically** significant or not?

4. Please answer the following questions about missig data.

(1) Why missing data in X is less problematic than missing data in Y?

(2) What is **censored data**? Please give an example.

(3) What is **truncated data**? Please give an example.

(4) What is **sample selection**? Please give an example.

(5) What is the difference between **censored data**, **truncated data** and **sample selection**?

4 Practical Exercise(I): Estimate the returns to schooling for Chinese Women(30 points)

Research Question: How education affect earnings for married women? Please refer to Zhang et al.(2005) for more details if you are interested in the research question.

- Please use **the data you had cleaned by yourself in HW1** to estimate the returns to schooling for Chinese women.
- Based on the data cleaning from HW1, you may need to make some adjustments. Restrict your sample to **married women** currently residing in urban areas including both rural-urban migrants and local residents. In this study, the variables you need to prepare include: *anualwage*, *educ*, *exper*, *provcode*, lfp(1 if in the labor market, 0 otherwise), *kids*(number of children). **Note:** After completing the cleaning process, the annualwage of women not in the labor force should be recorded as NA.

1. **OLS** for the return to education for working women

- Suppose we estimate the Mincer wage equation for women as follows:

$$lnY_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \gamma X_i + u_i$$

where

- lnY_i is the log of annual wage for individual i, X_i is a set of provincial dummy variables.

  i) Which parameter is the most of interest for our question? explain the economic meaning of it.

  ii) Please use the concept of internal validity to assessing our OLS estimation in the case (Hint: list potential threatens to the validity and explain them specifically)

2. **Nonlinear Regression**

- Our analysis reveals that a subset of women are out of the labor force; we will use models to predict their likelihood of participation. (Hint: the independent variables should include $educ$, $exper$, $exper^2$ $provcode$ and $kids$, the dependent variable is lfp)

  i) Please use the LPM model to estimate the probability of being in the labor force. What is the interpretation of the coefficient of $educ$? Is it a good model for this case? Why?

  ii) Please use the logit model to estimate the probability of being in the labor force. What is the interpretation of the coefficient of $educ$? what is the marginal effect of $educ$ on the probability of being in the labor force?

  iii) Please use the probit model to estimate the probability of being in the labor force. What is the interpretation of the coefficient of $educ$? what is the marginal effect of $educ$ on the probability of being in the labor force?

3. **Heckman Selection Model**

- To eliminate the potential bias, we use heckman selection model with 2 steps methods to estimate the regression again.

i) Explain why we may need heckman selection model to estimate the regression?

ii) Based on probit model in 2(ii), you can get the inverse Mills ratio. Please use the inverse Mills ratio to estimate the regression again. What is the interpretation of the coefficient of $educ$?[NOTE: following the better identification strategy, you have to add at least one more control variable than in outcome equation.]

iii) Reestimate the model using R Package(sampleSelection or others) to see if there a difference between the result here and above? If there is a difference, please explain why.

iv) Can you tell if there is evidence of sample selection bias? If so, how do you know? If not, how do you know?

4. **Bonus Question**

- As mentioned in 1(ii), the OLS estimation may be biased due to other factors. How can we control for the bias? Please use the method you think is appropriate to estimate the regression again.
- Please draw a DAG to demostrate and verify your choices. (Hint: How to draw DAG in R?, Please refer to the following link here)

5 Practical Exercise(II): Matching and Propensity Score Matching(30 points)

Research Question: Wage gap between urban natives and rural-urban migrants in China

- Data and Package: In this study, you are required to use the the data you have cleaned in `HW1` to estimate the wage gap between urban natives and rural-urban migrants in China using matching and propensity score matching methods.
- **Outcome**: hourly wage
- **Treatment**: urban native (treat=1) and rural-urban migrant (treat=0)
- **Covariates**: female, age,education,experience, work information, city/province dummy variables, etc.(you can choose other covariates based on the data)

1. Summarize the key variables and compare the means of covariates between urban natives (treat=1) and rural-urban migrants (treat=0). Evaluate the balance of covariates before matching. (Hint: you can just copy your results from question 5 in HW1,but please don't show the raw result in the terminal windows,but use `stargazer` or other packages to report the results)

2. Run a simple and multiple OLS linear regression models to estimate the wage gap between the urban natives and rural-urban migrants. And report the results of `hukou status`, explain the results. (Hint: you can just copy your results from question 5 in HW1,but please use `stargazer` or other packages to report the results)

3. Perform **Exact Matching** on all covariates you think are important. Compare the means of covariates between the treatment and control groups after matching with before matching. Do you think the matching was successful? Why or why not? (Hint: you can use `MatchIt` package to perform matching in R)

4. Perform **Mahalanobis Distance Matching** on all covariates you think are important. Compare the means of covariates between the treatment and control groups after matching with before matching. Do you think the matching was successful? Why or why not? (Hint: you can use `MatchIt` package to perform matching in R)

5. Estimate the **propensity score** using a logistic regression model. Plot propensity score densities for treated and control groups in one graph. And evaluate the **common support** after matching, like the figure in the lecture.

6. Perform **nearest neighbor matching with replacement** using the estimated propensity scores. Estimate the Average Treatment Effect on the Treated (ATT) and interpret the result. Check covariate balance after matching using standardized mean differences. (**Hint:** Use `love.plot()` function from `cobalt` package to visualize the balance of covariates before and after matching.)

7. Perform radius matching (caliper = 0.2) using the estimated propensity scores. Estimate the ATT and interpret the result. Check covariate balance before and after matching using standardized mean differences.

8. Perform IPW estimation to estimate the wage gap between the urban natives and rural-urban migrants.(Hint: you can use `ipw` package to estimate the IPW model in R)

9. In this study, compare the results of different methods and explain why matching methods are better than OLS.

6 Reference

Jia, R., Lan, X., & Miquel, G. P. i. (2021). Doing business in China: Parental background and government intervention determine who owns busines. Journal of Development Economics

Zhang et al.(2005) Economic returns to schooling in urban China, 1988 to 2001,Journal of Comparative Economics

This will be the last homework of this semester. Good Luck and Have a good holiday!