Introduction to Metrics: Homework 1

Business School, Nanjing University

Zhaopeng Qu

3/21/2025

目录

1 Learning Objectives

- Review key concepts from the lectures
- Develop familiarity with R and household survey datasets
- Learn and practice data cleaning techniques
- Understand the Rubin Causal Model and OLS regression analysis

2 Due Date and Submission Guidelines

- Due by April 7, 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

3.1 Mathematical Review (20 points):

- Based on content from the lectures, prove the following statements (*all notation follows that used in the lecture slides*):

1. If $D_i|X_i \perp (Y_{0i}, Y_{1i})$, then $E[Y_i|D_i = 1, X_i] = E[Y_i|X_i]$. Please use the continuous case in your proof.

2. Review the concept of conditional expectation to prove the **Law of Iterated Expectations** and its extensions for the discrete case:

$$E[Y_i] = E[E[Y_i|X_i]] \tag{3.1}$$

and

$$E[g(X_i)Y_i] = E\big[E[g(X_i)Y_i|X_i]\big] = E\big[g(X_i)E[Y_i|X_i]\big] \tag{3.2}$$

and in the special case

$$E[g(X_i)|X_i] = g(X_i) \tag{3.3}$$

3. Express the following concepts in mathematical form as presented in the lectures:

- the **sample mean**
- the **variance of the sample mean**
- the **sample covariance**
- the **standard error of the sample mean**
- the **standard error of the sample covariance**

and prove the following statements:

- the sample mean is an **unbiased** estimator of the population mean
- the sample covariance is a **consistent** estimator of the population covariance

3.2 Multiple Regression and Causal Inference (30 points):

1. Prove why **Omitted Variable Bias** leads to a **biased** estimator of β_1. (Hint: derive the formula for $\hat{\beta}_1$ in the OVB case, as shown in lecture 3 slide 30, then discuss briefly)

2. The Frisch-Waugh-Lovell theorem is an important theorem in econometrics that helps us understand how controlling for variables works in multiple regression. Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{3.4}$$

- Using the Frisch-Waugh-Lovell theorem, prove that the OLS estimator of β_1 can be obtained through a two-step partional regression of Y_i.
- In the context of causal inference, explain why "controlling" for X_2 and X_3 helps identify the causal effect of X_1 on Y. Use the potential outcomes framework to illustrate your point.

3. Use the `CASchools` data from the `AER` package to replicate the results shown in lecture 3 slides 54-59:

$$TestScore = \beta_0 + \beta_1 STR + \beta_2 lunch + u_i \tag{3.5}$$

- Please use the `CASchools` data from the `AER` package to replicate the results in lecture 3 slides 54-59. **Note** that the **control variable** is now `lunch` instead of `elpct`, so the regression equation becomes:

$$TestScore = \beta_0 + \beta_1 STR + \beta_2 \text{lunch} + u_i \tag{3.6}$$

- Verify that all summations of $\tilde{u}_i$, $\tilde{X}_i$, and their cross products with X_i equal zero, as shown on lecture 3 slides 49 and 55. (Hint: you can use the `lm` function in R to obtain the residuals and then calculate the summations)
- Report the estimated coefficients of `STR` using three methods:

1) the partial regression formula from lecture 3 slide 56

2) regression of Y on $\tilde{X}_i$ as in lecture 3 slide 57

3) regression of Y on X as in lecture 3 slide 58

4 Practical Exercise: Migrant-Native Wage Gap (50 points)

1. Obtaining household survey data

- Visit one of the following websites and create an account to access the data:
  - 中国居民收入调查,China Household Income Project(CHIP)
  - 中国家庭追踪调查,China Faimily Panel Studies(CFPS)
  - 中国健康与营养调查,China Health and Nutrition Survey(CHNS)
  - 中国健康养老追踪调查,China Health and Retirement Longitudinal Study(CHARLS)
  - 中国综合社会调查,Chinese General Social Survey(CGSS)
  - 中国劳动力动态调查,China Labor-force Dynamics Survey(CLDS)
  - 中国家庭金融调查,China Household Financial Survey(CHFS)

2. Data cleaning and preparation for summary statistics (30 points)

   - Select cross-sectional data from a specific year (e.g., CHIP2002, CHARLS2011, or another dataset)
   - Restrict your sample to households currently residing in urban areas including both rural-urban migrants and local residents.(In some datesets, you may need to combine the urban and rural data into one dataset, if the data is separated by urban household and migrant household).
   - Clean and prepare the following variables, then create a descriptive statistics table:

Variable Names	Descriptions
anualwage	年工资 = 月工资 *12+ 年终奖金以及其他货币化补贴
workhour	年工作小时数 = 每天工作小时 * 周或月工作天数 *12
workstatus	工作状况：工资/自我雇佣/失业/家务劳动等
female	= 1 if female; = 0 otherwise
age	年龄
educ	years of education (受教育年限)
hukou	户口状况：城市 ==1 or 农村 ==0
exper	working experience(if not available,can be calculate with age-educ-6
self-reported health(if available)	“excellent, very good, good ,fair and poor”and scale it from the number “1”to “5”
citycode(if available)	city code or name
provcode	province code or name

3. Descriptive Analysis (10 points)

    1) What are the average hourly wages for local residents and migrants in urban areas? Is there a statistically significant difference? Present your findings graphically (Hint: use grouped bar charts showing the differences, and include confidence intervals in your visualization)

    2) Are there significant differences between local residents and migrants across the other characteristic variables listed above? Present your findings in a table (Hint: structure the table with three columns: one for local resident statistics, one for migrant statistics, and one for the differences between them)

4. **Based on the chart and table above**, answer the following questions (30 points)

   1) Is there a wage gap between urban residents and rural migrants? How large is it? Provide at least two interpretations of the mean comparisons displayed in your chart. What are the possible underlying causal mechanisms?

   2) Suppose you want to test one of your interpretations. What specific mean comparisons would you make? Clearly identify your outcome variables and "treatment" variables.

   3) Following the Rubin Causal Model framework, formalize your question from part 2.

   4) Do your interpretations of the results have policy implications? If so, briefly discuss these implications. If not, briefly explain why your interpretations are insufficient to draw meaningful policy conclusions.

5. **Wage determination**

- A classical wage equation takes the form:

$$lnY_i = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_i \qquad (4.1)$$

where

- lnY_i represents the logarithm of **hourly wage** (or **annual wage** if working hours data is unavailable)

1) First, run the regression specified in equation (4.1):

   - Assuming that **education** and **experience** are your primary and secondary variables of interest, report the values of their estimated coefficients.

   - How can you be confident that these values represent true effects?

   - Interpret the meaning of β_1 in this regression.

- What is the return to an additional year of working experience?

2) Next, consider controlling for additional variables in the regression by estimating:

$$lnY_i = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + X'\gamma + u_i \qquad (4.2)$$

where X is a vector of demographic and other control variables such as gender, age, hukou status, health status, and location indicators.

- Add the corresponding variables **one by one** to the regression in the following order: city, gender, age and experience (only if the `experience` variable in your dataset is not directly computed as "age-educ-6"), hukou status, health status, etc.

- How does the estimated coefficient $\widehat{\beta}_1$ change as you add these controls? Provide your best explanation for why it changes in this manner.

3) Since the `hukou` variable indicates whether respondents are rural-urban migrants, report the value and statistical significance of its estimated coefficient in the full specification (including all variables in equation 4.2). Explain its economic implications.

4) If you remain interested in the effect of education on wages, how would you determine whether this effect differs significantly between urban natives and rural-urban migrants?

(Hint: you'd better put all regressions above into a table, then answer questions above in order)

Good Luck and Have Fun!