Homework 2

Introduction to Metrics,Fall 2023

Zhaopeng Qu

5/7/2023

目录

1 Learning Objectives

- Better understanding of the concept of Internal Validity, IV, Sample Selection methods and RDD

2 Due Date and Formats

- Due to May.21 24:00.
  - Late sending will lower your scores by a standard *"10% per hour deduction"*
- Required forms.

1. Upload your report(**inculding PDF or docx files**) and all raw files to generate it(such as `.stmd` and `.do` files for **Stata** users) to **教学立方** system. 关于该平台的使用可以参考 **使用指南：学生版**

2."All raw files" including *stmd*(`.stmd` for **Stata** users) as well as dofiles(`.do` for **Stata** users) to generate the final report,are requried to uploaded.

3. **Rename your files with** `ID_Name_Major_HW2` (eg. `151090001_`户纳东`_`经济`_HW2.zip`), other names or forms may not be accepted.

4.English or Chinese_(Writing in English will receive a small bonus)

3 Theory and Application Exercises

3.1 Exercises in SW textbooks(Third Edition,English version)(20 points)

- 7.1-7.6(pp284-286)
- 8.5 & 8.6(pp341-342)
- 9.11(pp385)
- 11.1-11.7(pp450-451)
- 12.2 & 12.3(pp498-499)
- 13.9(pp551)
- Please try to prove the following formula of IV estimator when there is population heterogeneity in the treatment effect and in the influence of the instrument on the receipt of treatment.

$$\hat{\beta}_{2SLS} \xrightarrow{p} \frac{Cov(ZY)}{Cov(ZX)} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}$$

3.2 Practical Exercise(I): Instrumental Variable and Heckman Selection Model(30 points)

Research Question: How education affect earnings for married women? Specifically, how much does a woman's earnings increase when she obtains an additional schooling year?

- The file mroz.dta from "T.A. Mroz (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," Econometrica 55, 765-799."

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

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

$$lnwage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i$$

1) Which parameter is the most of interest for our question? explain the economic meaning of it.
2) 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. **Instrumental Variable** to returns to education for working women

- To eliminate some potential biases, we use two variables *fatheduc* and *motheduc*, which are father's and mother's education as instruments for the own education.

1) Using the content you learned in the class to assessing whether *fatheduc* and *mothereduc* can be valid instruments for the estimation of return to education for working women or not. Which one (or both or neither)should be chosen as our instrument(s)?

2) Report the all regressions you run for your assessment in tables(one or two), which should include all results of

   (1) the Simple OLS regression
   (2) the OLS with covariates
   (3) the first stage regression of 2SLS
   (4) the reduced form regression
   (5) the second regression of 2SLS
   (6) Addtional regressions may be useful you think

Note that standard errors should be shown in parenthesis.

3) Comparing the estimation results of 2SLS and OLS, evaluate the true magnitude of the return to education for working women.

3. **Heckman Selection Model** to returns to education for working women

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

1) Explain why we may need heckman selection model to estimate the regression? Why instrumental variable method may not eliminate the bias from sample selection if it exists.
2) Estimate the model in two individual steps without exclusion restrictions, by predicting the inverse mills ratios and including this as a control variable in the wage equation.
3) Estimate the model in two individual steps with exclusion restrictions which includes the variables *nwifeinc*, *age*, *kidslt6*, and *kidsge6* in the selection model.
4) Reestimate two models using Stata command(*heckman*) or R Package(*sampleSelection* or others) to see if there a difference between the result here and above.
5) Can you tell if there is evidence of sample selection base on the last estimation above(thus using command to estimate it with exclusion restrictions)

4. **Bonus Question**(10 points)

- How to estimate the return to education for working women taking both endogenous variable of education and sample selection? (Hint: which means we have to combine IV and sample selection methods to estimate the regression)

3.3 Practical Exercise(II): Replicates results using urban Chinese data(20 points)

Research Question: Estimate the returns to schooling for **women** in the Chinese labor market

- Please try to use **the data you had cleaned by yourself in HW1** to *replicate the main results in the Exercise(I)* to estimate the returns to schooling for Chinese women. (Hint: you might have to choose some different covariates, instrumental variables and variables in exclusion restrictions from **mroz.dta**)
- Reference: 黄志岭和姚先国,"教育回报率的性别差异研究", 《世界经济》,2009 年第 7 期。

3.4 Practical Exercise(III) RDD Estimation and Specifications(30 points)

- **Reference**: Christopher Carpenter and Carlos Dobkin(2009),"The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age",Am Econ J Appl Econ. 2009 January 1; 1(1): 164–182.
- Data and Variable Descriptions: The data file mlda.dta contains data the National Center for Health Statistics (NCHS) confidential mortality detail files for 1997-2004. These data are derived from death certificates and cover all deaths in the United States in the study period.The sample is restricted to fatalities of young adults aged 19-22. The data used here consist of averages in 48 cells defined by age in 30-day intervals.
- Causes of death are divided into *internal* and *external*, with the latter

split into mutually exclusive subcategories: homicide, suicide, motor vehicle accidents, and other external causes. A separate category for alcohol-related causes covers all deaths for which alcohol was mentioned on the death certificate.

- Outcomes are mortality rates per 100,000, where the denominator comes from census population estimates.

Variables	Description
all	Death rate from all causes(per 100,000)
internal	Death rate from internal causes (per 100,000)
external	Death rate from external causes (per 100,000)
alcohol	Death rate from alcohol-related causes (per 100,000)
homicide	Death rate from homicide causes (per 100,000)
suicide	Death rate from suicide causes (per 100,000)
mva	Death rate from motor vehicle accidents (per 100,000)
drugs	Death rate from drugs-related causes (per 100,000)
externalother	Death rate from other external causes (per 100,000)
agecell	age in month

1) Which is the running variable in the case? Center it around 21 and call it *X_c*. And generate the treatment variable and call it *D*.

2) Plot the relationship between the running variables and the outcome using different specifications, including **linear functions with the same and different slopes** on both sides of the cutoff and **quadratic functions**. Can you identify any discontinuities at the threshold in the graphs?

3) Write down the specific equations used and report the estimated results for both **linear with different slopes and quadratic functions**. Provide a detailed explanation of your findings.

4) Next,after limiting your samples to only those between 20 and 22 years old, rerun the regressions for both linear functions with different slopes and quadratic functions. Provide a detailed explanation of your updated findings. Is there any difference from the previous results?

5) Last,restore your sample to those between 19-23 years old, and replace your outcomes with *internal*, *external*, *MVA*, and *alcohol*. Then, rerun the regressions for **linear functions with different slopes and quadratic functions**. Provide a detailed explanation of your findings. What conclusions can you draw?