Instrumental Variable Estimation 2: Implementation in R

Instructor: Yuta Toyama

Last updated: 2020-03-30

Section 1

Introduction

Introduction

- I cover three examples of instrumental variable regressions.
  1. Wage regression
  2. Demand curve
  3. Effects of Voter Turnout

Section 2

Wage regression

Example 1: Wage regression

- Use dataset "Mroz", cross-sectional labor force participation data that accompany "Introductory Econometrics" by Wooldridge.
  - Original data from *"The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions"* by Thomas Mroz published in *Econometrica* in 1987.
  - Detailed description of data: https://www.rdocumentation.org/packages/npsf/versions/0.4.2/topics/mroz

```
library("foreign")

# You might get a message "cannot read factor labels from Stat
data <- read.dta("MROZ.DTA")
```

```
## Warning in read.dta("MROZ.DTA"): cannot read factor labels
```

- Describe data

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression

##  R package version 5.2.2. https://CRAN.R-project.org/packag
```

```r
stargazer(data, type = "text")
```

```
##
## ==================================================================
## Statistic  N     Mean     St. Dev.   Min   Pctl(25) Pctl(75
## ------------------------------------------------------------------
## inlf      753   0.568     0.496      0      0         1
## hours     753  740.576   871.314     0      0       1,516
## kidslt6   753   0.238     0.524      0      0         0
```

- Consider the wage regression

$$\log(w_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \epsilon_i$$

- We assume that $exper_i$ is exogenous but $educ_i$ is endogenous.
- As an instrument for $educ_i$, we use the years of schooling for his or her father and mother, which we call $fathereduc_i$ and $mothereduc_i$.
- Discussion on these IVs will be later.

```r
library("AER")
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

- How about the first stage? You should always check this!!

```
# First stage regression

result_1st <- lm(educ ~ motheduc + fatheduc + exper + expersq,

# F test
linearHypothesis(result_1st,
                 c("fatheduc = 0", "motheduc = 0" ),
                 vcov = vcovHC, type = "HC1")
```

```
## # A tibble: 2 x 4
##   Res.Df    Df     F  `Pr(>F)`
##    <dbl> <dbl> <dbl>     <dbl>
## 1    425    NA  NA   NA
## 2    423     2  48.6  9.67e-20
```

Discussion on IV

- Labor economists have used family background variables as IVs for education.
- Relevance: OK from the first stage regression.
- Independence: A bit suspicious. Parents' education would be correlated with child's ability through quality of nurturing at an early age.
- Still, we can see that these IVs can mitigate (though may not eliminate completely) the omitted variable bias.
- Discussion on the validity of instruments is crucial in empirical research.

Section 3

Demand curve

Example 2: Estimation of the Demand for Cigaretts

- Demand model is a building block in many branches of Economics.
- For example, health economics is concerned with the study of how health-affecting behavior of individuals is influenced by the health-care system and regulation policy.
- Smoking is a prominent example as it is related to many illnesses and negative externalities.
- It is plausible that cigarette consumption can be reduced by taxing cigarettes more heavily.
- Question: how much taxes must be increased to reach a certain reduction in cigarette consumption? -> Need to know **price elasticity of demand** for cigaretts.

- Use `CigarrettesSW` in the package `AER`.
- a panel data set that contains observations on cigarette consumption and several economic indicators for all 48 continental federal states of the U.S. from 1985 to 1995.
- What is **panel data**? The data involves both time series and cross-sectional information.
    - The variable is denoted as y_{it}, which indexed by individual i and time t.
    - Cross section data y_i: information for a particular individual i (e.g., income for person i).
    - Time series data y_t: information for a particular time period (e.g., GDP in year y)
    - Panel data y_{it}: income of person i in year t.
- We will see more on panel data later in this course. For now, we use the panel data as just cross-sectional data (**pooled cross-sections**)

```r
# load the data set and get an overview
library(AER)
data("CigarettesSW")
summary(CigarettesSW)
```

```
##      state       year          cpi          population
##  AL     : 2   1985:48   Min.   :1.076   Min.   :  478447
##  AR     : 2   1995:48   1st Qu.:1.076   1st Qu.: 1622606
##  AZ     : 2             Median :1.300   Median : 3697472
##  CA     : 2             Mean   :1.300   Mean   : 5168866
##  CO     : 2             3rd Qu.:1.524   3rd Qu.: 5901500
##  CT     : 2             Max.   :1.524   Max.   :31493524
##  (Other):84
##      income               tax            price          
##  Min.   :  6887097   Min.   :18.00   Min.   : 84.97   Min.
##  1st Qu.: 25520384   1st Qu.:31.00   1st Qu.:102.71   1st Q
##  Median : 61661644   Median :37.00   Median :137.72   Media
##  Mean   : 99878736   Mean   :42.68   Mean   :143.45   Mean
```

- Consider the following model

$$\log(Q_{it}) = \beta_0 + \beta_1 \log(P_{it}) + \beta_2 \log(income_{it}) + u_{it}$$

  where

  - Q_{it} is the number of packs per capita in state i in year t,
  - P_{it} is the after-tax average real price per pack of cigarettes, and
  - $income_{it}$ is the real income per capita. This is demand shifter.
- As an IV for the price, we use the followings:
  - $SalesTax_{it}$: the proportion of taxes on cigarettes arising from the general sales tax.
    - Relevant as it is included in the after-tax price
    - Exogenous(indepndent) since the sales tax does not influence demand directly, but indirectly through the price.
  - $CigTax_{it}$: the cigarett-specific taxes

```r
library(dplyr)
CigarettesSW %>%
  mutate( rincome = (income / population) / cpi) %>%
  mutate( rprice  = price / cpi ) %>%
  mutate( salestax = (taxs - tax) / cpi ) %>%
  mutate( cigtax = tax/cpi ) -> Cigdata
```

- Let's run the regressions

```
cig_ols <- lm(log(packs) ~ log(rprice) + log(rincome) ,  data
#coeftest(cig_ols, vcov = vcovHC, type = "HC1")



cig_ivreg <- ivreg(log(packs) ~ log(rprice) + log(rincome)  |
                   log(rincome) +  salestax +  cigtax, data =
#coeftest(cig_ivreg, vcov = vcovHC, type = "HC1")

# Robust standard errors
rob_se <- list(sqrt(diag(vcovHC(cig_ols, type = "HC1"))),
               sqrt(diag(vcovHC(cig_ivreg, type = "HC1"))))

# Show result
stargazer(cig_ols, cig_ivreg, type ="text", se = rob_se)
```

```
##
## ==========================================================
```

- The first stage regression

```
# First stage regression

result_1st <- lm(log(rprice) ~ log(rincome) + log(rincome) + s

# F test
linearHypothesis(result_1st,
                 c("salestax = 0", "cigtax = 0" ),
                 vcov = vcovHC, type = "HC1")
```

```
## # A tibble: 2 x 4
##   Res.Df    Df     F  `Pr(>F)`
##    <dbl> <dbl> <dbl>     <dbl>
## 1     94    NA    NA  NA
## 2     92     2  128.   2.81e-27
```

Section 4

Voting

Example 3: Effects of Turnout on Partisan Voting

- THOMAS G. HANSFORD and BRAD T. GOMEZ "Estimating the Electoral Effects of Voter Turnout" The American Political Science Review Vol. 104, No. 2 (May 2010), pp. 268-288
  - Link: https://www.cambridge.org/core/journals/american-political-science-review/article/estimating-the-electoral-effects-of-voter-turnout/8A880C28E79BE770A5CA1A9BB6CF933C
- Here, we will see a simplified version of their analysis.
- The dataset is here

```r
library(readr)

HGdata <- read_csv("HansfordGomez_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )

## See spec(...) for full column specifications.
```

```r
stargazer::stargazer(as.data.frame(HGdata), type="text")
```

```
##
## ==================================================================
## Statistic             N       Mean       St. Dev.      Min
## ------------------------------------------------------------------
## Year               27,401  1,973.972     16.111       1,948
## FIPS_County        27,401 29,985.500   13,081.250     4,001
```

- Data description:

Name	Description
Year	Election Year
FIPS_County	FIPS County Code
Turnout	Turnout as Pcnt VAP
Closing2	Days between registration closing date and election
Literacy	Literacy Test
PollTax	Poll Tax
Motor	Motor Voter
GubElection	Gubernatorial Election in State
SenElection	U.S. Senate Election in State
GOP_Inc	Republican Incumbent
Yr52	1952 Dummy
Yr56	1956 Dummy
Yr60	1960 Dummy
Yr64	1964 Dummy
Yr68	1968 Dummy
Yr72	1972 Dummy

- Consider the following regression

$$DemoShare_{it} = \beta_0 + \beta_1 Turnout_{it} + u_t + u_{it}$$

  where

  - $Demoshare_{it}$: Two-party vote share for Democrat candidate in county i in the presidential election in year t
  - $Turnout_{it}$: Turnout rate in county i in the presidential election in year t
  - u_t: **Year fixed effects**. Time dummies for each presidential election year
- As an IV, we use the rainfall measure denoted by `DNormPrcp_KRIG`

```
# You can do this, but it is tedious.
hg_ols <- lm( DemVoteShare2 ~ Turnout + Yr52 + Yr56 + Yr60 + Y
                Yr72 + Yr76 + Yr80 + Yr84 + Yr88 + Yr92 + Yr96
#coeftest(hg_ols, vcov = vcovHC, type = "HC1")

# By using "factor(Year)" as an explanatory variable, the regr
hg_ols <- lm( DemVoteShare2 ~ Turnout + factor(Year)    ,  data
#coeftest(hg_ols, vcov = vcovHC, type = "HC1")

# Iv regression
hg_ivreg <- ivreg( DemVoteShare2 ~ Turnout + factor(Year) |
                    factor(Year) + DNormPrcp_KRIG, data = HGda
#coeftest(hg_ivreg, vcov = vcovHC, type = "HC1")

# Robust standard errors
rob_se <- list(sqrt(diag(vcovHC(hg_ols, type = "HC1"))),
               sqrt(diag(vcovHC(hg_ivreg, type = "HC1"))))
```