Instrumental Variable Estimation 2: Implementation in R

Instructor: Yuta Toyama

Last updated: 2021-05-18

Introduction

Introduction

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

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 do not have to worry about a message "cannot read factor labels from Stata 5 files".
data <- read.dta("data/MROZ.DTA")
```

```
## Warning in read.dta("data/MROZ.DTA"): cannot read factor labels from Stata 5
## files
```

- Describe data

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

```
##
## ===============================================================
## Statistic  N      Mean      St. Dev.    Min   Pctl(25) Pctl(75)  Max
## ---------------------------------------------------------------
## inlf       753    0.568       0.496      0       0        1       1
## hours      753   740.576    871.314      0       0      1,516   4,950
## kidslt6    753    0.238       0.524      0       0        0       3
## kidsge6    753    1.353       1.320      0       0        2       8
## age        753    42.538      8.073      30      36       49      60
## educ       753    12.287      2.280      5       12       13      17
## wage       428    4.178       3.310    0.128   2.263    4.971   25.000
## repwage    753    1.850       2.420    0.000   0.000    3.580   9.980
## hushrs     753 2,267.271    595.567     175    1,928    2,553   5,010
## husage     753    45.121      8.059      30      38       52      60
## huseduc    753    12.491      3.021      3       11       15      17
## huswage    753    7.482       4.231    0.412   4.788    9.167   40.509
## faminc     753 23,080.600 12,190.200 1,500   15,428   28,200  96,000
## mtr        753    0.679       0.083    0.442   0.622    0.721   0.942
## motheduc   753    9.251       3.367      0       7        12      17
## fatheduc   753    8.809       3.572      0       7        12      17
## unem       753    8.624       3.115      3      7.5       11      14
## city       753    0.643       0.480      0       0        1       1
## exper      753    10.631      8.069      0       4        15      45
## nwifeinc   753    20.129      11.635   -0.029  13.025   24.466  96.000
## lwage      428    1.190       0.723    -2.054  0.817    1.604   3.219
## expersq    753   178.039    249.631      0       16       225   2,025
## ---------------------------------------------------------------
```

- 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")
library("dplyr")
library("texreg")
library("estimatr")


# data cleaning
data %>%
  select(lwage, educ, exper, expersq, motheduc, fatheduc) %>%
  filter( is.na(lwage) == 0 ) -> data


result_OLS <- lm_robust( lwage ~ educ + exper + expersq, data = data, se_type = "HC1")

# IV regression using fathereduc and mothereduc
result_IV <- iv_robust(lwage ~ educ + exper + expersq |
                         fatheduc + motheduc + exper + expersq,
                       data = data, se_type = "HC1")

# Show result
screenreg(l = list(result_OLS, result_IV),digits = 3,
          # caption = 'title',
         # custom.model.names = c("(I)", "(II)", "(III)", "(IV)", "(V)"),
          custom.coef.names = NULL,  # add a class, if you want to change the names of variables.
          include.ci = F,include.rsquared = FALSE, include.adjrs = TRUE, include.nobs = TRUE,
          include.pvalues = FALSE, include.df = FALSE, include.rmse = FALSE,
          custom.header = list("lwage" = 1:2), # you can add header especially to indicate dependen
          stars = numeric(0))
```

```
##
## ==============================
##                       lwage
##              ------------------
##              Model 1   Model 2
## ------------------------------
## (Intercept)   -0.522     0.048
##               (0.202)   (0.430)
## educ           0.107     0.061
##               (0.013)   (0.033)
## exper          0.042     0.044
##               (0.015)   (0.016)
## expersq       -0.001    -0.001
##               (0.000)   (0.000)
## ------------------------------
## Adj. R^2       0.151     0.130
## Num. obs.    428       428
## ==============================
```

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

```
# First stage regression

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

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

```
## Linear hypothesis test
##
## Hypothesis:
## fatheduc = 0
## motheduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ motheduc + fatheduc + exper + expersq
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    425
## 2    423  2 48.644 < 2.2e-16 ***
## ---
```

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.

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
data("CigarettesSW")
summary(CigarettesSW)
```

```
##      state          year          cpi          population           packs       
##  AL     : 2   1985:48   Min.   :1.076   Min.   :  478447   Min.   : 49.27  
##  AR     : 2   1995:48   1st Qu.:1.076   1st Qu.: 1622606   1st Qu.: 92.45  
##  AZ     : 2             Median :1.300   Median : 3697472   Median :110.16  
##  CA     : 2             Mean   :1.300   Mean   : 5168866   Mean   :109.18  
##  CO     : 2             3rd Qu.:1.524   3rd Qu.: 5901500   3rd Qu.:123.52  
##  CT     : 2             Max.   :1.524   Max.   :31493524   Max.   :197.99  
##  (Other):84                                                                 
##      income                tax             price             taxs       
##  Min.   :  6887097   Min.   :18.00   Min.   : 84.97   Min.   : 21.27  
##  1st Qu.: 25520384   1st Qu.:31.00   1st Qu.:102.71   1st Qu.: 34.77  
##  Median : 61661644   Median :37.00   Median :137.72   Median : 41.05  
##  Mean   : 99878736   Mean   :42.68   Mean   :143.45   Mean   : 48.33  
##  3rd Qu.:127313964   3rd Qu.:50.88   3rd Qu.:176.15   3rd Qu.: 59.48  
##  Max.   :771470144   Max.   :99.00   Max.   :240.85   Max.   :112.63  
##
```

- 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

```
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_robust(log(packs) ~ log(rprice) + log(rincome) ,  data = Cigdata,se_type = "HC1")
#coeftest(cig_ols, vcov = vcovHC, type = "HC1")


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

# Show result
screenreg(l = list(cig_ols, cig_ivreg),digits = 3,
          # caption = 'title',
          custom.model.names = c("OLS", "IV"),custom.coef.names = NULL,  # add a class, if you want
          include.ci = F,include.rsquared = FALSE, include.adjrs = TRUE, include.nobs = TRUE,
          include.pvalues = FALSE, include.df = FALSE, include.rmse = FALSE,
          custom.header = list("log(packs)" = 1:2), # you can add header especially to indicate dep
          stars = numeric(0)
          )
```

```
##
## ==============================
##                   log(packs)
##                ---------------
##                OLS      IV
## ------------------------------
## (Intercept)   10.067    9.736
##               (0.502)  (0.514)
## log(rprice)   -1.334   -1.229
##               (0.154)  (0.155)
## log(rincome)   0.318    0.257
##               (0.154)  (0.153)
## ------------------------------
## Adj. R^2       0.542    0.539
## Num. obs.     96       96
## ==============================
```

- The first stage regression

```
# First stage regression

result_1st <- lm(log(rprice) ~ log(rincome) + log(rincome) + salestax + cigtax , data= Cigdata)

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

```
## Linear hypothesis test
##
## Hypothesis:
## salestax = 0
## cigtax = 0
##
## Model 1: restricted model
## Model 2: log(rprice) ~ log(rincome) + log(rincome) + salestax + cigtax
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     94
## 2     92  2 127.77 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
library(readr)

HGdata <- read_csv("data/HansfordGomez_Data.csv")

stargazer::stargazer(as.data.frame(HGdata) %>% select(-starts_with("Yr")),type="text")
```

```
##
## ==========================================================================
## Statistic            N       Mean     St. Dev.    Min   Pctl(25)  Pctl(75)     Max
## --------------------------------------------------------------------------
## Year              27,401 1,973.972    16.111    1,948    1,960     1,988     2,000
## FIPS_County       27,401 29,985.500 13,081.250  4,001   20,013    39,157    56,045
## Turnout           27,401   65.562     10.514   20.366   58.477    72.613    100.000
## Closing2          27,401   23.053     13.042      0       11        30        125
## Literacy          27,401   0.058      0.234       0       0         0         1
## PollTax           27,401   0.001      0.023       0       0         0         1
## Motor             27,401   0.211      0.408       0       0         0         1
## GubElection       27,401   0.434      0.496       0       0         1         1
## SenElection       27,401   0.680      0.467       0       0         1         1
## GOP_Inc           27,401   0.501      0.500       0       0         1         1
## DNormPrcp_KRIG    27,401   0.005      0.208    -0.419   -0.093     0.001     2.627
## GOPIT             27,401   33.282     34.066      0       0       66.3        100
## DemVoteShare2_3MA 27,401   44.250     10.606   10.145   37.006    50.996    88.982
## DemVoteShare2     27,401   43.622     12.415    6.420   34.954    51.858    97.669
## RainGOPI          27,401   0.007      0.142      -0      -0.03       0         2
## TO_DVS23MA        27,401 2,886.877   792.530   473.161 2,321.025 3,384.772 8,526.616
## Rain_DVS23MA      27,401   0.355      10.188   -25.054  -4.019     0.028    144.257
## dph               27,401   0.021      0.145       0       0         0         1
## dvph              27,401   0.018      0.133       0       0         0         1
## rph               27,401   0.025      0.155       0       0         0         1
## rvph              27,401   0.025      0.155       0       0         0         1
## state_del         27,401   0.037      0.187    -0.821   -0.090     0.172     0.619
## --------------------------------------------------------------------------
```

```
##
## ================================================================================
## Statistic         N       Mean        St. Dev.      Min   Pctl(25)  Pctl(75)     Max
## --------------------------------------------------------------------------------
## dph_StateVAP   27,401 77,525.150   597,474.000     0         0         0     6,150,988
## dvph_StateVAP  27,401 63,138.400   663,707.600     0         0         0     12,700,000
## rph_StateVAP   27,401 243,707.900 1,720,659.000    0         0         0     18,300,000
## rvph_StateVAP  27,401 142,166.500 1,071,445.000    0         0         0     12,800,000
## State_DVS_lag  27,401   46.896        8.317     22.035   40.767    52.197     80.872
## State_DVS_lag2 27,401  2,268.381     786.199    485.533 1,661.934 2,724.515 6,540.244
## --------------------------------------------------------------------------------
```

- 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

Name	Description
Yr52	1952 Dummy
Yr56	1956 Dummy
Yr60	1960 Dummy
Yr64	1964 Dummy
Yr68	1968 Dummy
Yr72	1972 Dummy
Yr76	1976 Dummy
Yr80	1980 Dummy
Yr84	1984 Dummy
Yr88	1988 Dummy
Yr92	1992 Dummy
Yr96	1996 Dummy
Yr2000	2000 Dummy

Name	Description
DNormPrcp_KRIG	Election day rainfall - differenced from normal rain for the day
GOPIT	Turnout x Republican Incumbent
DemVoteShare2_3MA	Partisan composition measure = 3 election moving avg. of Dem Vote Share
DemVoteShare2	Democratic Pres Candidate's Vote Share
RainGOPI	Rainfall measure x Republican Incumbent
TO_DVS23MA	Turnout x Partisan Composition measure
Rain_DVS23MA	Rainfall measure x Partisan composition measure
dph	=1 if home state of Dem pres candidate
dvph	=1 if home state of Dem vice pres candidate

Name	Description
rph	=1 if home state of Rep pres candidate
rvph	=1 if home state of Rep vice pres candidate
state_del	avg common space score for the House delegation
dph_StateVAP	= dph*State voting age population
dvph_StateVAP	= dvph*State voting age population
rph_StateVAP	= rph*State voting age population
rvph_StateVAP	= rvph*State voting age population
State_DVS_lag	State-wide Dem vote share, lagged one election
State_DVS_lag2	State_DVS_lag squared

- 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`

```r
# You can do this, but it is tedious.
hg_ols <- lm_robust( DemVoteShare2 ~ Turnout + Yr52 + Yr56 + Yr60 + Yr64 + Yr68 + Yr72 + Yr76 + Yr8(
                     + Yr84 + Yr88 + Yr92 + Yr96 + Yr2000,  data = HGdata, se_type="HC1")
#coeftest(hg_ols, vcov = vcovHC, type = "HC1")

# By using "factor(Year)" as an explanatory variable, the regression automatically incorporates the
hg_ols <- lm_robust( DemVoteShare2 ~ Turnout + factor(Year) , data = HGdata, se_type="HC1")
#coeftest(hg_ols, vcov = vcovHC, type = "HC1")

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

# Show result
screenreg(l = list(hg_ols, hg_ivreg),
          digits = 3,
          # caption = 'title',
          custom.model.names = c("OLS", "IV"),
          custom.coef.names = NULL,  # add a class, if you want to change the names of variables.
          include.ci = F,
          include.rsquared = FALSE, include.adjrs = TRUE, include.nobs = TRUE,
          include.pvalues = FALSE, include.df = FALSE, include.rmse = FALSE,
          custom.header = list("DemVoteShare2" = 1:2), # you can add header especially to indicate (
          stars = numeric(0)
          )
```

```
##
## ====================================
##                   DemVoteShare2
##               ---------------------
##               OLS          IV
## ------------------------------------
## (Intercept)     59.085       26.910
##                 (0.560)     (11.024)
## Turnout         -0.157        0.363
##                 (0.008)      (0.178)
## ------------------------------------
## Adj. R^2         0.280        0.130
## Num. obs.    27401        27401
## ====================================
```

```r
# First stage regression
hg_1st <- lm(Turnout ~  factor(Year) + DNormPrcp_KRIG, data= HGdata)

# F test
linearHypothesis(hg_1st,
                 c("DNormPrcp_KRIG = 0" ),
                 vcov = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## DNormPrcp_KRIG = 0
##
## Model 1: restricted model
## Model 2: Turnout ~ factor(Year) + DNormPrcp_KRIG
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  27387
## 2  27386  1 44.029 3.296e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```