Intro to instrumental variable regression in R

Malcolm Gillies

7 December 2022

main@01f1ed9 2022-12-05

Acknowledgement

The material in this presentation was adapted from:

Introduction to Econometrics with R
Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer
2021-10-06
`https://www.econometrics-with-r.org/`

Especially Chapter 12 / Section 12.1

main@01f1ed92022-12-05

Note on terminology

- Econometrics and (bio)statistics use different words for things
- A glossary can help, e.g. Gunasekara et al. [2008]
- *Endogenous regressor* ~ Confounded independent variable

main@01f1ed9 2022-12-05

IV assumptions

1. Strongly related to the independent variable
2. Uncorrelated with confounders and only related to outcome via the independent variable

main@01f1ed92022-12-05

Notation

- X_i independent variable
- Z_i instrumental variable
- Y_i outcome variable

main@01f1ed9 2022-12-05

Two-stage least-squares regression

First stage: regress independent variable on IV and predict "good part" of independent variable

$$X_i = \underbrace{\pi_0 + \pi_1 Z_i}_{\text{good}} + \underbrace{\nu_i}_{\text{bad}} \tag{1}$$

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i \tag{2}$$

Second stage: regress outcome on predictions to yield effect β_1

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i \tag{3}$$

Example: how do cigarette sales vary with price?

- Does increasing the price of cigarettes cause a decrease in sales?
  - *Elasticity of demand*
- Not confounding per se
  - *Simultaneous causality* (two-way causality) in supply and demand
- Instrumental variable analysis can remove bias

main@01f1ed9 2022-12-05

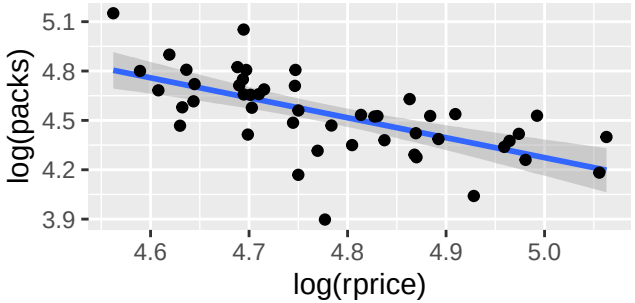
Setup

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

#### compute real per capita prices
CigarettesSW$rprice <- with(CigarettesSW, price / cpi)

####  compute the sales tax
CigarettesSW$salestax <- with(CigarettesSW, (taxs - tax) / cpi)

#### generate a subset for the year 1995
c1995 <- subset(CigarettesSW, year == "1995")
```

main@01f1ed9 2022-12-05

Scatterplot

```
library(ggplot2)
ggplot(data=c1995, aes(x=log(rprice), y=log(packs))) +
    geom_smooth(method='lm', formula= y~x) + geom_point()
```



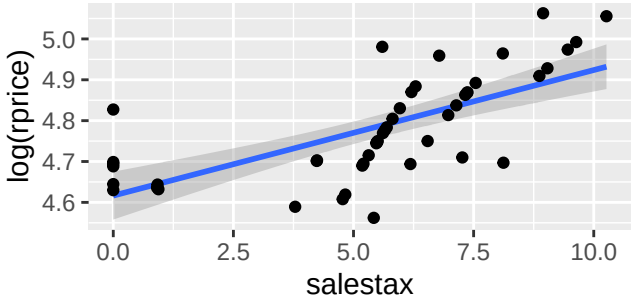
main@01f1ed9 2022-12-05

Naive estimate

```
summary(lm(log(packs)~log(rprice), data=c1995))$coefficients

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.338924  1.0352902  9.986499 4.248876e-13
## log(rprice) -1.213057  0.2164497 -5.604336 1.129667e-06
```

Scatterplot (instrumental variable)

```
library(ggplot2)
ggplot(data=c1995, aes(x=salestax, y=log(rprice))) +
    geom_smooth(method='lm', formula= y~x) + geom_point()
```



main@01f1ed9 2022-12-05

First stage regression (1)

```
cig_s1 <- lm(log(rprice) ~ salestax, data = c1995)

coeftest(cig_s1, vcov = vcovHC, type = "HC1")

##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 4.6165463  0.0289177 159.6444 < 2.2e-16 ***
## salestax    0.0307289  0.0048354   6.3549 8.489e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

main@01f1ed9 2022-12-05

First stage regression (2)

```
#### inspect the R^2 of the first stage regression
summary(cig_s1)$r.squared

## [1] 0.4709961

#### store the predicted values
lcigp_pred <- cig_s1$fitted.values
```

main@01f1ed9 2022-12-05

Second stage regression

```
cig_s2 <- lm(log(c1995$packs) ~ lcigp_pred)
coeftest(cig_s2, vcov = vcovHC)

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  9.71988    1.70304  5.7074 7.932e-07 ***
## lcigp_pred  -1.08359    0.35563 -3.0469  0.003822 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

main@01f1ed9 2022-12-05

Automated TSLS using `ivreg()`

```
cig_ivreg <- ivreg(log(packs) ~ log(rprice) | salestax, data = c1995)
coeftest(cig_ivreg, vcov = vcovHC, type = "HC1")

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  9.71988    1.52832  6.3598 8.346e-08 ***
## log(rprice) -1.08359    0.31892 -3.3977  0.001411 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check strength of instrument using F test

```
linearHypothesis(cig_s1, "salestax=0", vcov = vcovHC, type = "HC1")

## Linear hypothesis test
##
## Hypothesis:
## salestax = 0
##
## Model 1: restricted model
## Model 2: log(rprice) ~ salestax
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     47
## 2     46  1 35.713 3.145e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

main@01f1ed9 2022-12-05

Additional worked examples

- *Causal Inference in Education*
- `https://bookdown.org/aschmi11/causal_inf/instrumental-variables-estimation.html`
- A tutorial on the use of instrumental variables in pharamcoepidemiology Ertefaie et al. [2017]

main@01f1ed9 2022-12-05

References

A. Ertefaie, D. S. Small, J. H. Flory, and S. Hennessy. A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety*, 26(4):357–367, 2017. ISSN 1099-1557. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.4158.

F. I. Gunasekara, K. Carter, and T. Blakely. Glossary for econometrics and epidemiology. *Journal of Epidemiology and Community Health (1979-)*, 62(10):858–861, 2008. URL https://www.jstor.org/stable/20720836.

main@01f1ed9 2022-12-05