Econometrics 1 *Applied Econometrics with R*

Lecture 11: Instrumental Variable

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼1726

E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Office hour: Mon./Tue. 13:00-14:00

Omitted variable bias

- Definition:

  If a regressor is correlated with a variable that has been omitted from the analysis and that determines, in part, the dependent variable, then the OLS estimator of the corresponding coefficient will have omitted variable bias.

- An example:

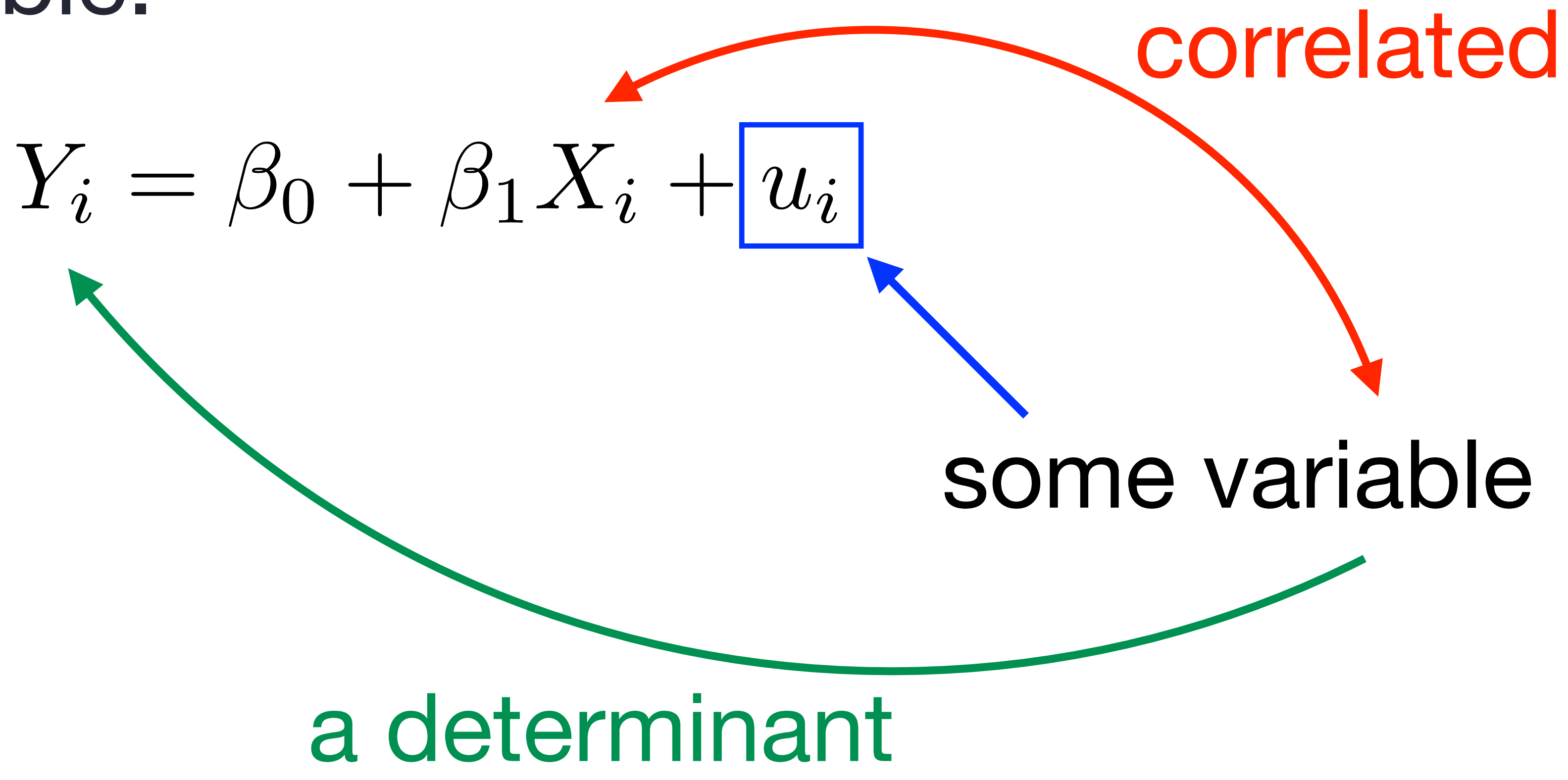
  Dependent variable — the test score
  Regressor — the student-teacher ratio
  Omitted variable — the percentage of English learners

Omitted variable bias

- Omitted variable bias occurs when:

  1. the omitted variable is correlated with the included regressor, and

  2. the omitted variable is a determinant of the dependent variable.



$$Y_i = \beta_0 + \beta_1 X_i + \boxed{u_i}$$

correlated

some variable

a determinant

Omitted variable bias

- Assumption 1 of OLS

$$\mathrm{E}(u_i \mid X_i) = 0$$

- If X and u are correlated, this assumption is violated and the OLS estimator is biased.

- Let the correlation between X_i and u_i be

$$\mathrm{corr}(X_i, u_i) = \rho_{Xu}$$

then,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

meaning that the OLS estimator is inconsistent.

The “Mozart effect”

- A study published in *Nature* in 1993 suggested that listening to Mozart for 10 to 15 minutes could temporarily raise your IQ by 8 to 9 points.
  Rauscher, Shaw, and Ky, Music and spatial task performance, *Nature*, vol.365, pp. 611, 1993.

- Evidence of “Mozart effect”?
  Many studies found that students who take optional music or arts courses in high school have higher English and math test scores.

- There is omitted variable bias in those studies.

How to address omitted variable bias?

- The simple way:

  Include the omitted variable in a multiple regression, if you have data on the omitted variable.

- The general way:

  Instrumental variable regression

Instrumental variable

$$Y_i = \beta_0 + \beta_1 \boxed{X_i} + u_i$$

- Suppose X has two parts:

  1) the part correlated with u, and
  2) the part uncorrelated with u

- We need to find a way to isolate the second part.

Instrumental variable

$$Y_i = \beta_0 + \beta_1 \boxed{X_i} + u_i$$

- X and u are correlated. A variable Z is an “instrumental variable”, or “instrument”, if it satisfies

1. Instrument relevance: $\mathrm{corr}(Z_i, X_i) \neq 0$, and

2. Instrument exogeneity: $\mathrm{corr}(Z_i, u_i) = 0$

The two stage least squares (TSLS) estimator

- First stage — run regression between X and Z using OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- Second stage — regress Y with the predicted $\hat{X}$ using OLS

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

$$Y_i = \beta_0^{\text{TSLS}} + \beta_1^{\text{TSLS}} \hat{X}_i + u_i^{\text{TSLS}}$$

The TSLS estimator

- The TSLS estimator $\hat{\beta}_1^{\mathrm{TSLS}}$ is consistent and normally distributed in large samples, but still biased.

- When there is a single regressor X and a single IV Z,

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{ZY}}{s_{ZX}}$$

then, when the sample size increases,

$$\hat{\beta}_1^{\mathrm{TSLS}} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)} = \beta_1$$

The Philip Wright’s problem

- How to set an import tariff (tax) on animal and vegetable oils and fats, such as butter and soy oil.

- The key to understand the economic effect of a tariff was having quantitative estimates of the demand and supply curves of the goods.

- The demand equation:

$$\ln(Q_i^{\text{butter}}) = \beta_0 + \underset{\underset{\text{demand elasticity}}{\uparrow}}{\beta_1} \ln(P_i^{\text{butter}}) + u_i$$

The Philip Wright's problem

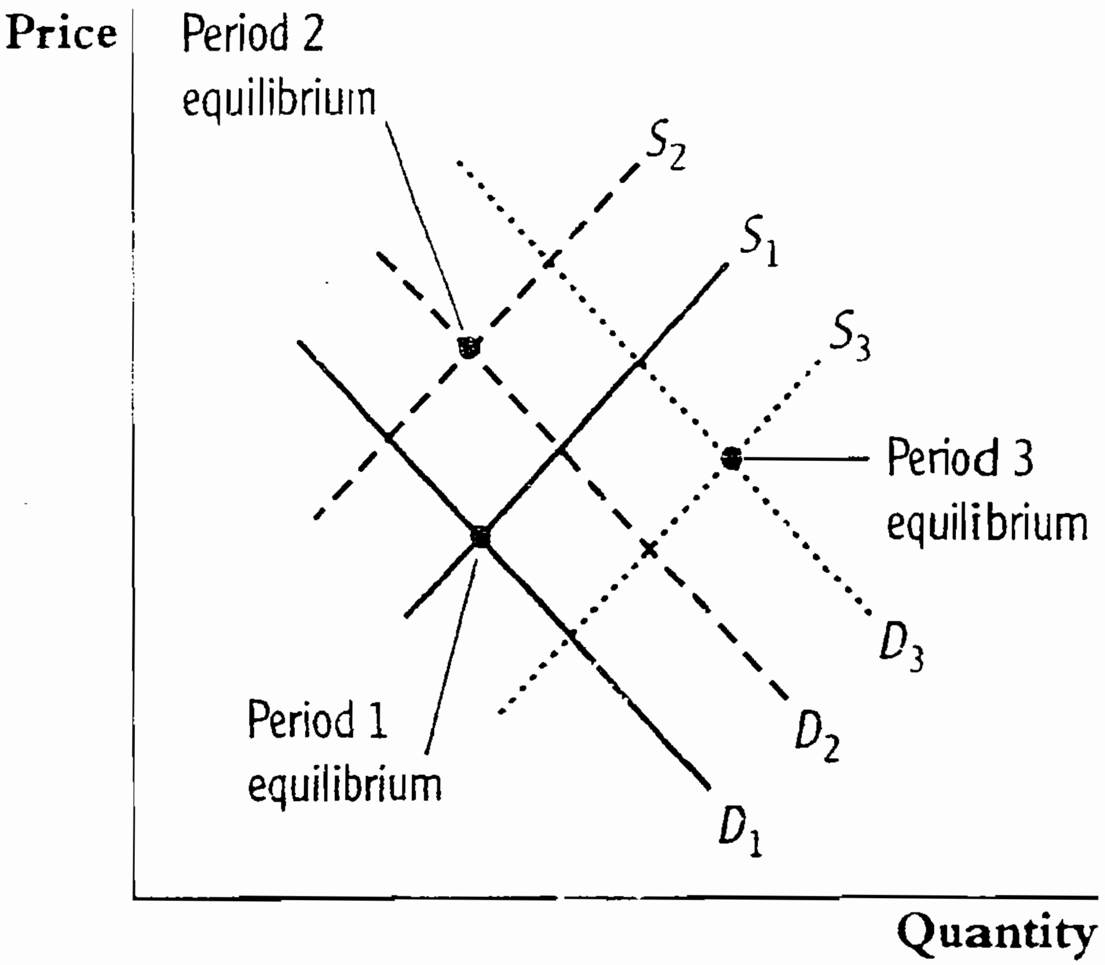
- Philip Wright had data on total annual butter consumption and its average annual price in the US for 1912 through 1922.

- With the data it would be easy to estimate the demand elasticity by OLS.

- However, because of the interaction between demand and supply, the regressor was likely to be correlated with the error term.

The interaction between demand and supply

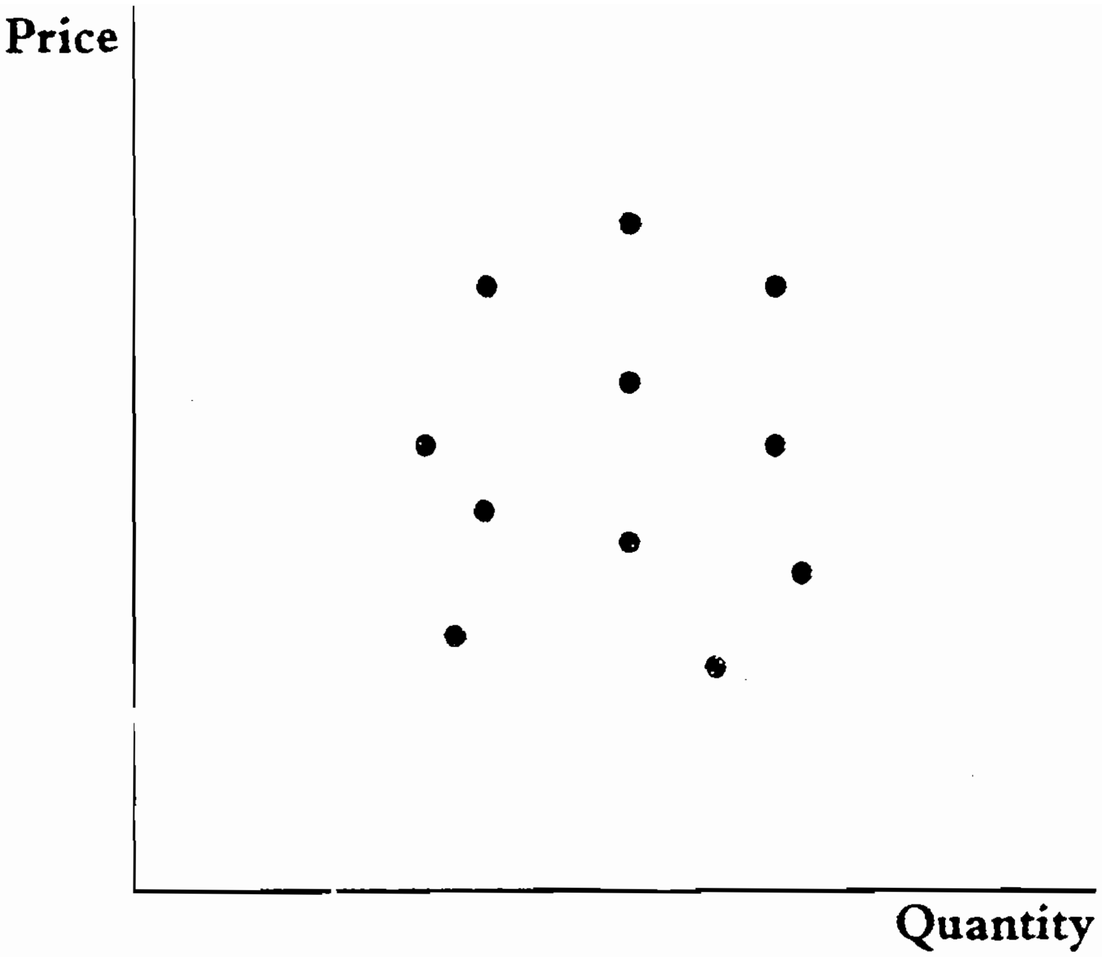
(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the third period is the intersection of D_3 and S_3.



(a) Demand and supply in three time periods

The interaction between demand and supply

(b) This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



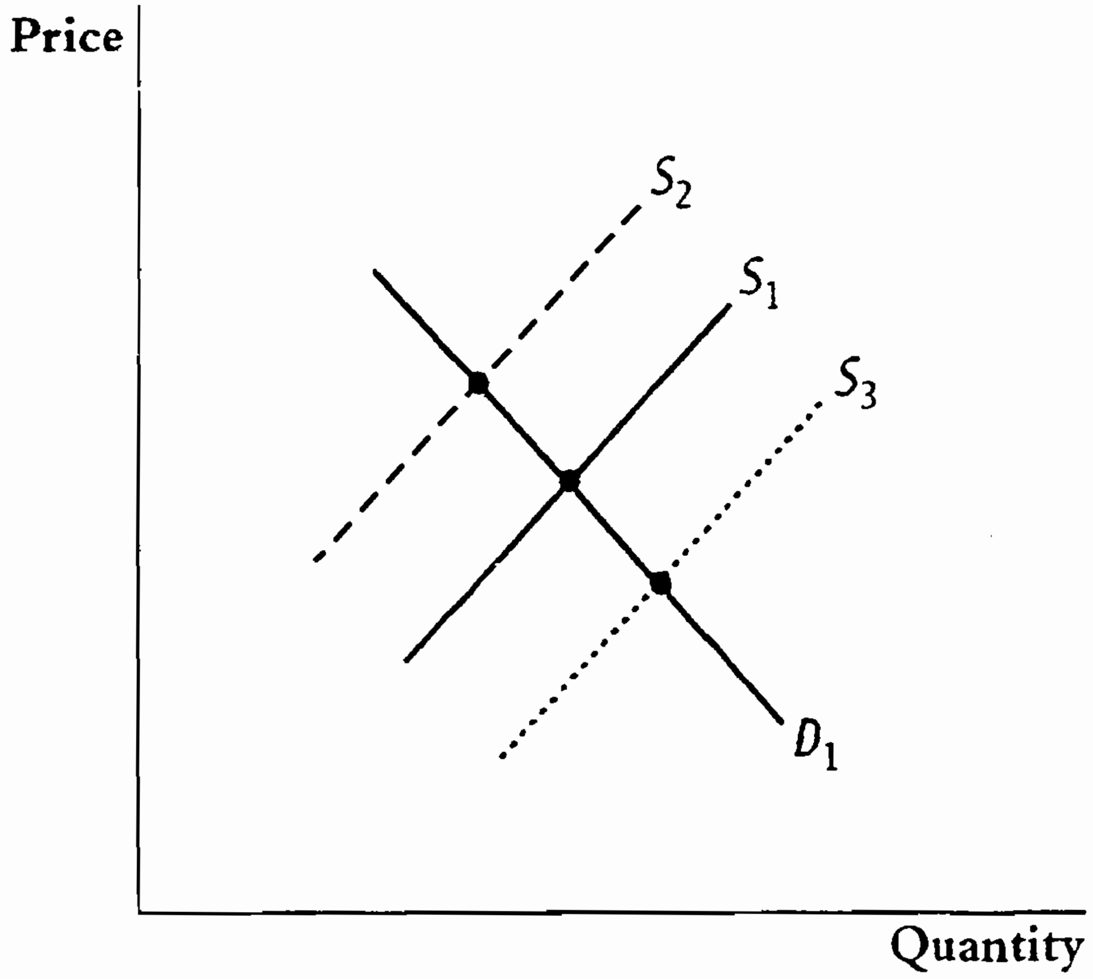
(b) Equilibrium price and quantity for 11 time periods

The interaction between demand and supply

- Wright realized that a way to get around this problem was to find some third variable that shifted supply but did not shift demand.

The interaction between demand and supply

(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1, the equilibrium prices and quantities trace out the demand curve.



(c) Equilibrium price and quantity when only the supply curve shifts

The interaction between demand and supply

$$\ln(Q_i^{\text{butter}}) = \beta_0 + \beta_1 \ln(P_i^{\text{butter}}) + u_i$$

- In the IV formulation, the third variable (the IV) is correlated with the price (it shifts the supply curve, which leads to a change in price) but is uncorrelated with u (the demand curve remains stable).

- One potential IV is the weather.

Application to the demand for cigarettes

- What would the after-tax sales price of cigarettes need to be to achieve 20% reduction in cigarette consumption?

- To answer this question, we need to know the elasticity of demand for cigarettes. If the elasticity is -1, then the 20% target in consumption can be achieved by a 20% increase in price. That is to say, the sales tax must be 20%.

- We don't know the elasticity and must estimate it from data on prices and sales.

The data “CigarettesSW” in the AER package

- The data “CigarettesSW” in the AER package contains the data of cigarette consumption of the 48 continental US States for 1985 and 1995.

- There are 7 variables other than `state` and `year`:

`cpi`	Consumer price index.
`population`	State population.
`packs`	Number of packs per capita.
`income`	State personal income (total, nominal).
`tax`	Ave. state, federal, and ave. local excise taxes for fiscal year.
`price`	Average price during fiscal year, including sales tax.
`taxs`	Average excise taxes for fiscal year, including sales tax.

The TSLS estimation

- The original regression

$$\ln(\underbrace{Q_i^{\text{cigarettes}}}_{\text{No. of packs per capita in the state}}) = \beta_0 + \beta_1 \ln(\underbrace{P_i^{\text{cigarettes}}}_{\text{ave. price per pack}}) + u_i$$

- First stage of TSLS regression

$$\ln(P_i^{\text{cigarettes}}) = \pi_0 + \pi_1 \underbrace{SalesTax_i}_{\text{calculated as } (\texttt{taxs - tax})} + v_i$$

- Second stage of TSLS regression

$$\ln(Q_i^{\text{cigarettes}}) = \beta_0^{\text{TSLS}} + \beta_1^{\text{TSLS}} \ln(\widehat{P_i^{\text{cigarettes}}}) + u_i^{\text{TSLS}}$$

Practice with the 1995 data

- Data preparation

```
> library("AER")
> data("CigarettesSW")

> c1995 <- subset(CigarettesSW, year ==
"1995")
> packs <- c1995$packs
> rprice <- c1995$price / c1995$cpi
> salestax <- (c1995$taxs - c1995$tax) /
c1995$cpi
```

Practice with the 1995 data

- Perform the OLS estimation of

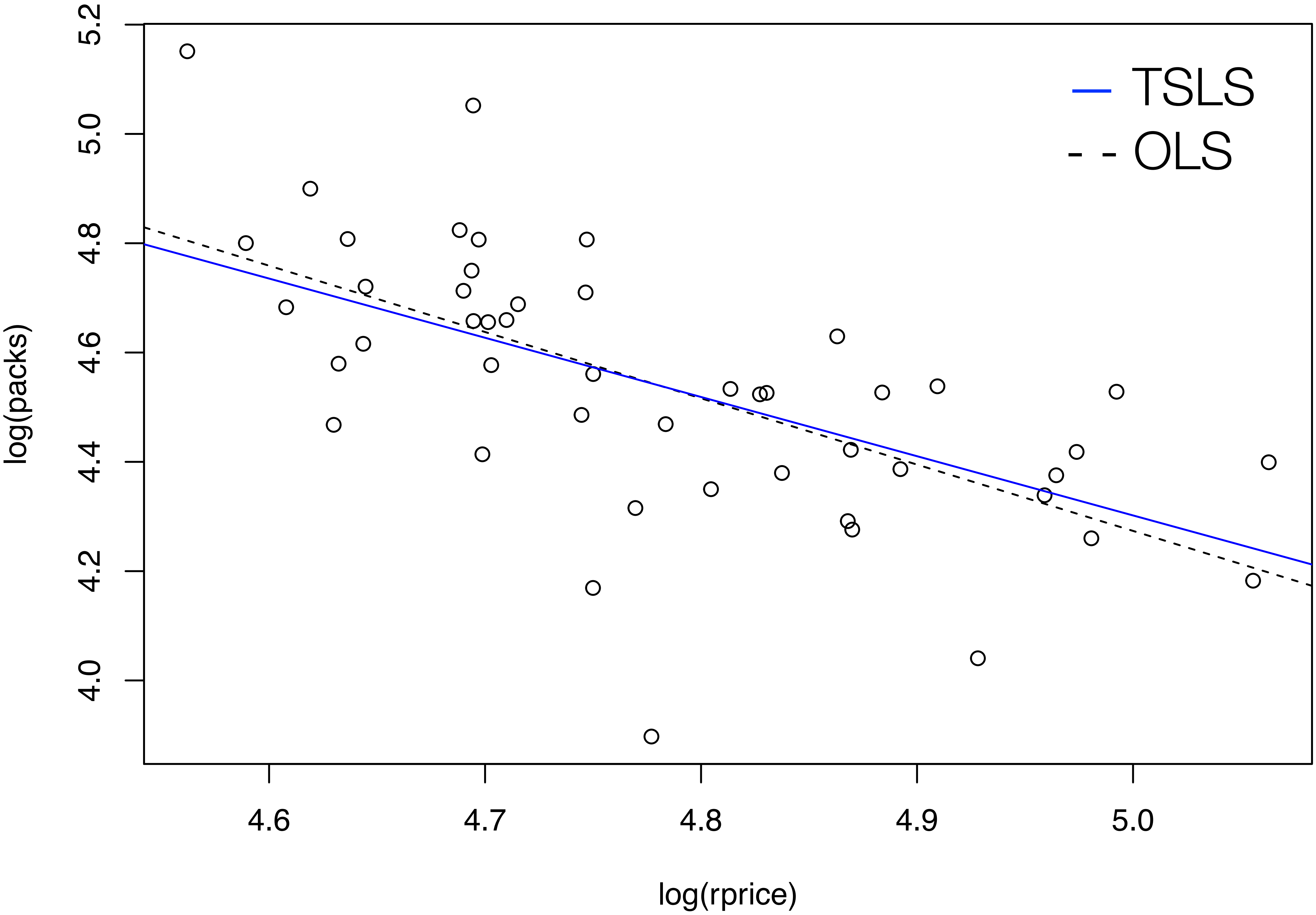
$$\ln(Q_i^{\text{cigarettes}}) = \beta_0 + \beta_1 \ln(P_i^{\text{cigarettes}}) + u_i$$

- Perform the TSLS estimation of the IV regression

$$\ln(P_i^{\text{cigarettes}}) = \pi_0 + \pi_1 SalesTax_i + v_i$$

$$\ln(Q_i^{\text{cigarettes}}) = \beta_0^{\text{TSLS}} + \beta_1^{\text{TSLS}} \widehat{\ln(P_i^{\text{cigarettes}}}) + u_i^{\text{TSLS}}$$

- Compare the two estimates $\hat{\beta}_1$ and $\hat{\beta}_1^{\text{TSLS}}$


TSLS
OLS
log(packs)
log(rprice)
4.0
4.2
4.4
4.6
4.8
5.0
5.2
4.6
4.7
4.8
4.9
5.0

The `ivreg` command

- As we have seen, the TSLS method can be performed using the OLS method (`lm` command).

- There is a single command `ivreg` who can do the same thing.

```
> ivreg(y ~ x | z)
```

endogenous regressors

instrumental variables

```
> ivreg(y ~ x1 + x2 | z1 + z2 + z3)
```

- Apply `ivreg` to the IV regression of cigarettes.

`lm()` vs. `ivreg()`

- Compare the results of the previous IV model generated by applying `lm()` and `ivreg()` commands.

- What do you think is the most important difference?

`lm()` vs. `ivreg()`

- Compare the results of the previous IV model generated by applying `lm()` and `ivreg()` commands.

- What do you think is the most important difference?

- The standard errors of coefficients on the second stage are not correct in `lm()` outputs.

The general IV regression model

- The general IV regression model has four types of variables:

  the dependent variable, Y,
  problematic *endogenous* regressors, X,
  included *exogenous* variables, W, and
  instrumental variables, Z.

- In general, there can be several X’s, W’s, and Z’s.

- The number of Z’s must be as least as many as the number of X’s.

The general IV regression model

- The general IV model:

$$\begin{aligned} Y = \beta_0 &+ \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \\ &+ \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i \end{aligned}$$

where

$X_{1i}, \ldots, X_{ki}$ are potentially correlated with u_i ,

$W_{1i}, \ldots, W_{ri}$ are uncorrelated with u_i , and

$Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

TSLS in the general IV model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \\ &\quad + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i \end{aligned}$$

- First stage estimation

$$\begin{aligned} X_{1i} &= \pi_{1,0} + \pi_{1,1} Z_{1i} + \cdots + \pi_{1,m} Z_{mi} \\ &\quad + \pi_{1,m+1} W_{1i} + \cdots + \pi_{1,m+r} W_{ri} + v_{1,i} \\ &\vdots \\ X_{ki} &= \pi_{k,0} + \pi_{k,1} Z_{1i} + \cdots + \pi_{k,m} Z_{mi} \\ &\quad + \pi_{k,m+1} W_{1i} + \cdots + \pi_{k,m+r} W_{ri} + v_{k,i} \end{aligned}$$

The `ivreg` command with included exogenous variables

-
  ```
  > ivreg(y ~ x + w1 + w2 | w1 + w2 + z1 + z2)
  ```

included exogenous variables
on both sides of “|”

Practice

- Take the logarithm of real income per capita as an included exogenous variable (i.e., W).

```
> rincome <- c1995$income /
c1995$population / c1995$cpi
```

- Take the sales tax (`salestax`) and the cigarettes specified tax (`tax` column in the data) as two IVs.

```
> cigtax <- c1995$tax / c1995$cpi
```

Practice

- Perform the IV regression using `ivreg` command to the following model

$$\ln(Q_i^{\text{cigarettes}}) = \beta_0 + \beta_1 \ln(P_i^{\text{cigarettes}}) + \beta_2 \ln(Income_i) + u_i$$

where $\ln(Income_i)$ is included exogenous variable, and

$$SalesTax_i, CigTax_i$$

are two instrumental variables.

The IV regression assumptions

1. $\mathrm{E}(u_i \mid W_{1i}, \ldots, W_{ri}) = 0$;

2. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution;

3. Large outliers are unlikely;

4. (1) Instrument Relevance

   (2) Instrument Exogeneity

The validity of IV (advanced)

- Whether IV regression is useful depends on whether the instrumental variables are valid.
- **Instrument relevance** — the instrumental variables must explain much of the endogenous regressors.

  If IVs explain little of the variation in X, they are called *weak instruments*.

  Weak instruments leads to biased TSLS estimator and unreliable *t*-statistics and confidence intervals.

The validity of IV (advanced)

- Weak instruments — a rule of thumb

  When there is a single endogenous regressor, then the first stage *F*-statistic can be a measure for checking for weak instruments.

  If the first stage *F*-statistic is less than 10, then the instruments are weak.

The validity of IV (advanced)

- **Instrument exogeneity**

  If the instruments are correlated to error terms, then the IV regression will not provide a consistent estimator.

- The judgement relies on expert knowledge when the number of X's equals the number of Z's.

- There is a statistical tool, called the *J*-statistic, can help when the number of X's is less than the number of Z's.

Where do valid instruments come from?

- In practice, the most difficult aspect of IV estimation is finding instruments that are both relevant and exogenous.
- There are two main approaches:
  - To use economic theory to suggest instruments
  - To use expert knowledge of the problem being studied, and careful attention to the details of data
- Suggestion → read papers, discuss with others.

References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.

2. Kleiber, C. and Zeileis, A., *Applied Econometrics with R*, Springer, 2008.