Econometrics 1

Lecture 12: Instrumental Variables (1)

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: `huangjp@szu.edu.cn`

Tel: (0755) 2695 0548

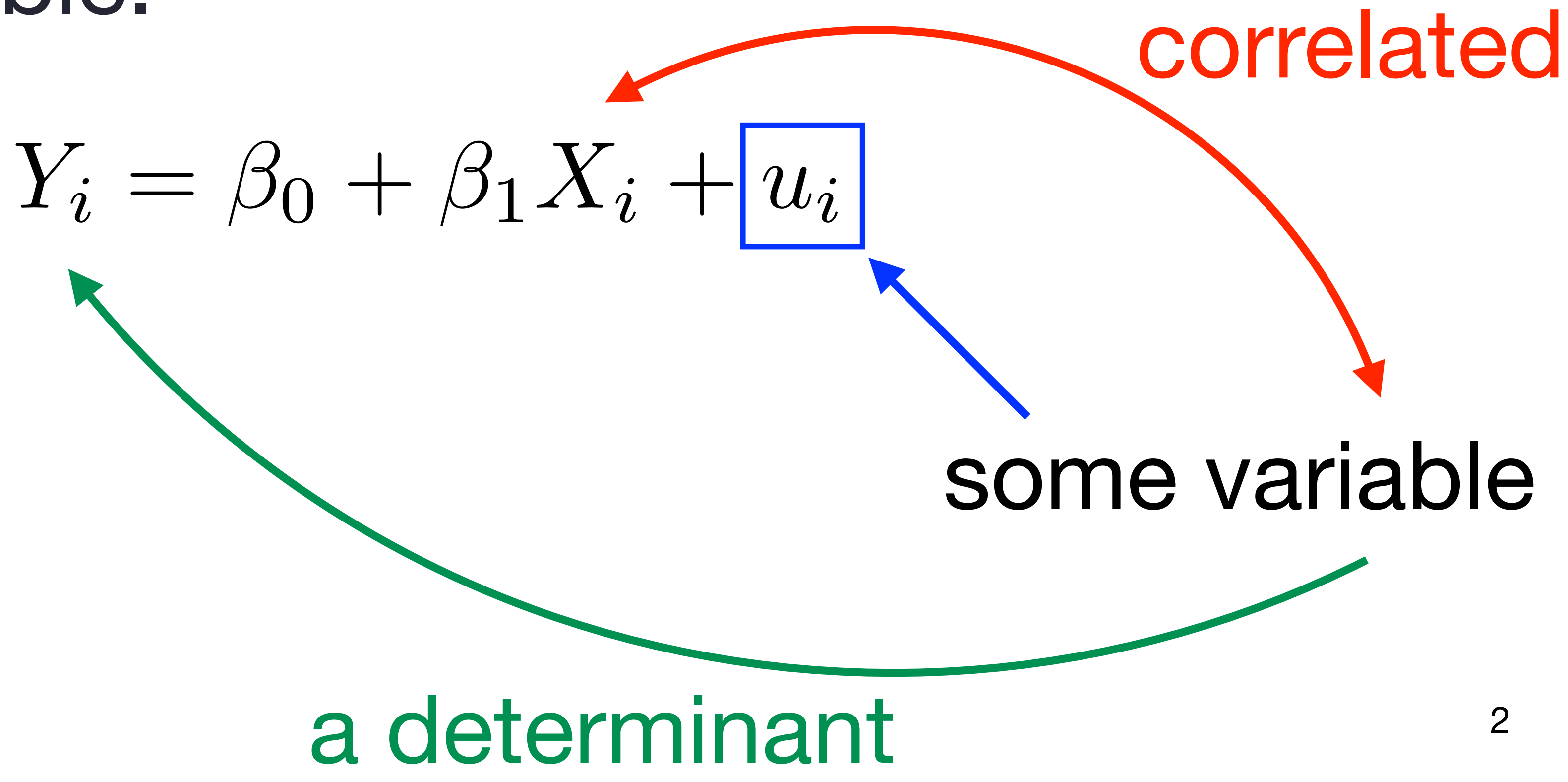
Website: `https://huangjp.com`

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 biased and 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.

- A choice when the simple way is not available:

  Instrumental variable regression

Instrumental variable

correlated

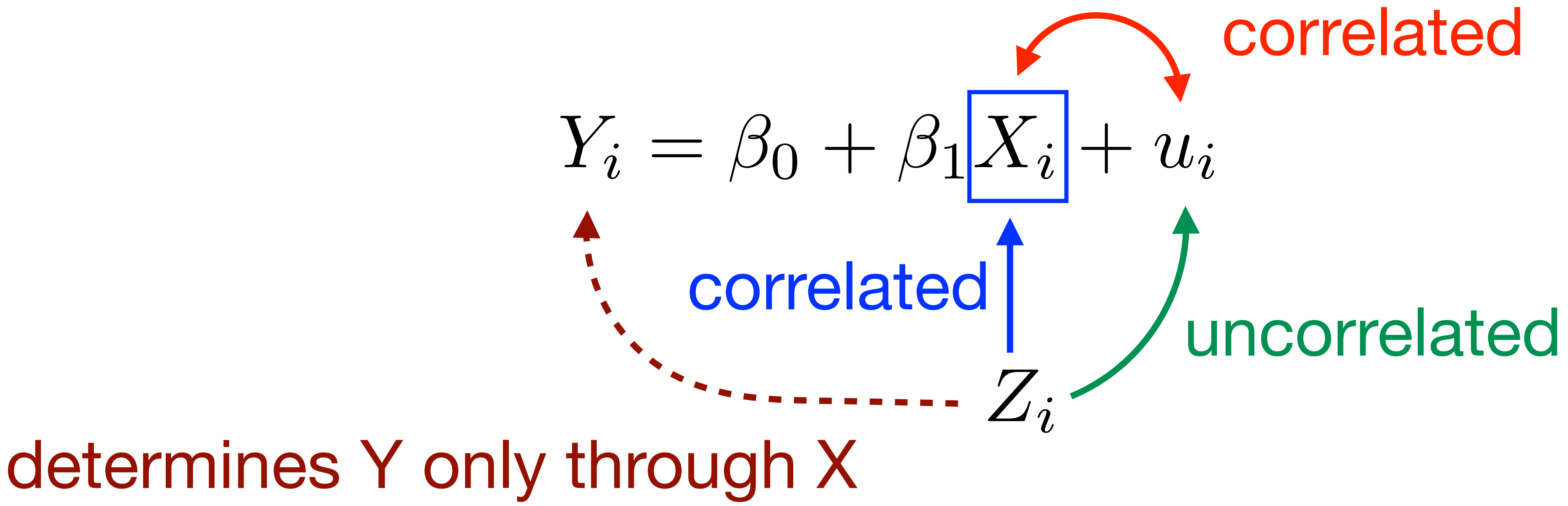
$$Y_i = \beta_0 + \beta_1 \boxed{X_i} + u_i$$

- 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



correlated

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

correlated

uncorrelated

Z_i

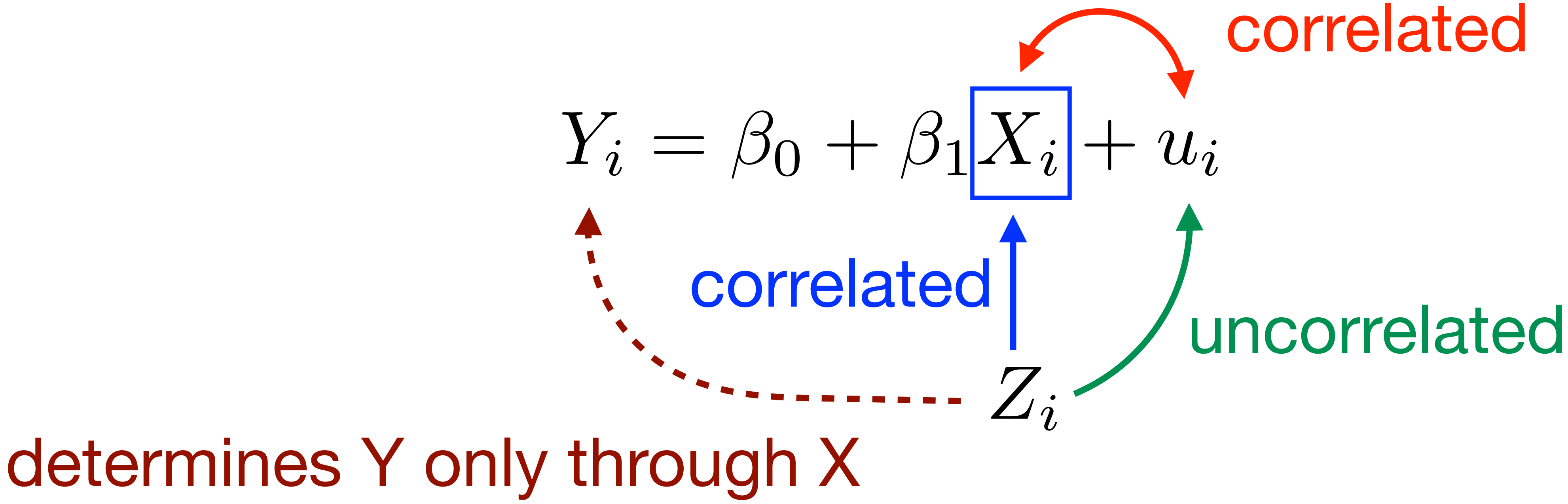
determines Y only through X

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

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

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

The IV estimator



correlated

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

correlated

uncorrelated

Z_i

determines Y only through X

- The IV estimator is

$$\beta_1^{\mathrm{IV}} = \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)}$$

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, for large samples,

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

$$\hat{\beta}_1^{\mathrm{TSLS}} \sim N\left(\beta_1^{\mathrm{IV}}, \frac{1}{n}\frac{\mathrm{var}[(Z_i - \mu_Z)\mu_i]}{[\mathrm{cov}(Z_i, X_i)]^2}\right)$$

When to use IV regression?

- The IV regression can be used when X and u are correlated (X is **endogenous**), where the OLS estimation is biased.

- There are several typical situations, including:
  1. Omitted variable bias (unobserved);
  2. Simultaneous causality;
  3. Measurement errors.

Simultaneous causality

- Simultaneous causality means both X causes Y and Y causes X. For example, student teacher ratio and test score.

- Simultaneous equations

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

If u_i is negative which decreases Y_i, then the value of X_i through the second equation is lowered if γ_1 is X_i and u_i are positively correlated.

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:

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

The Philip Wright's problem

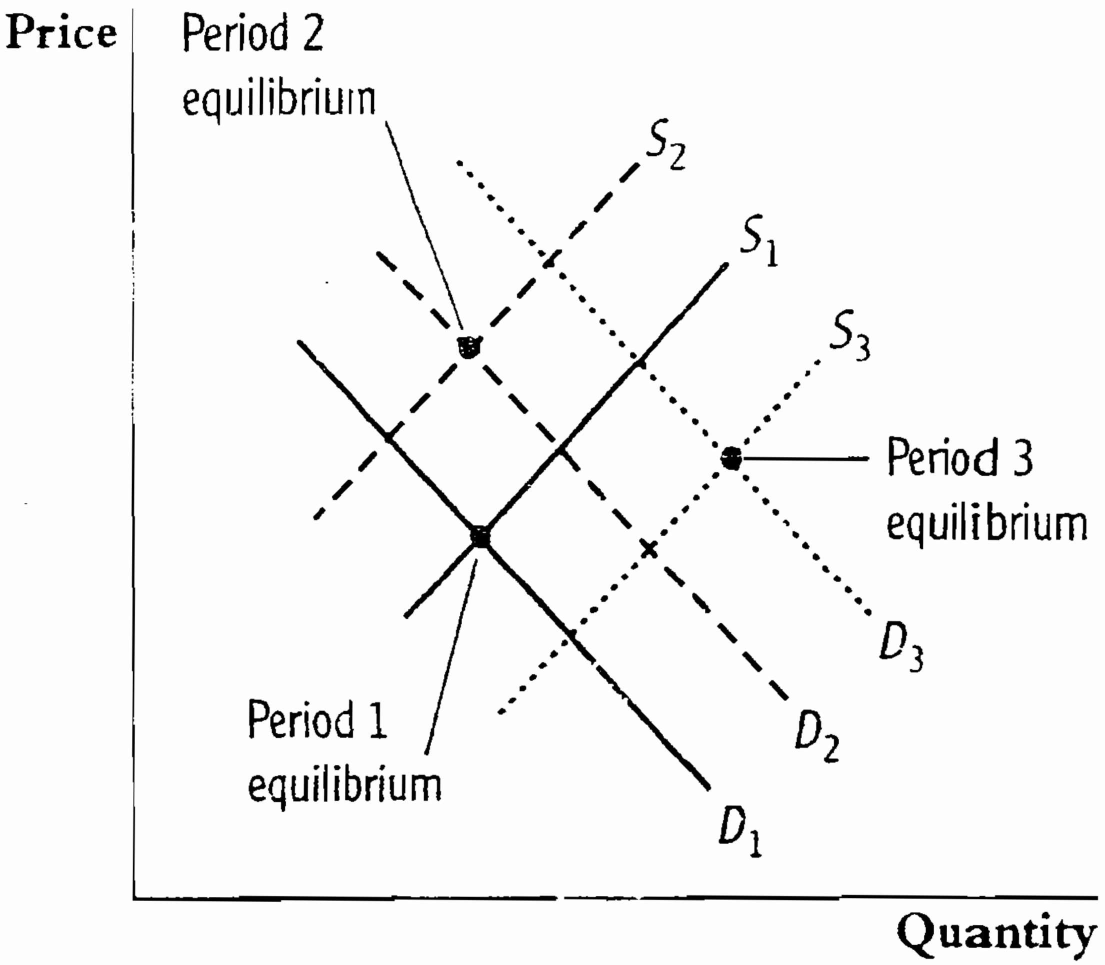
- Philip Wright had data on total annual butter consumption and its average annual price in the US for 1912 — 1922.

- 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

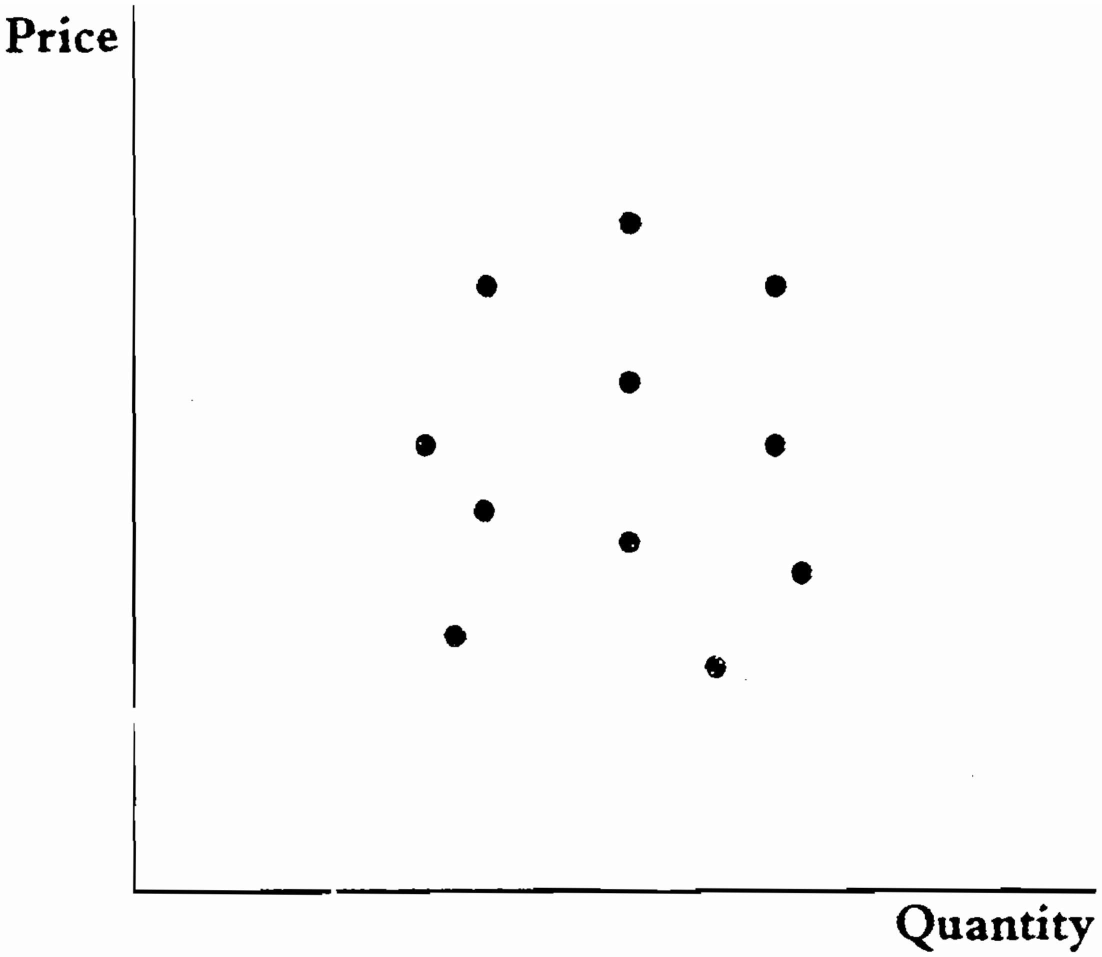
(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve D_1 and the supply curve S_1. Equilibrium in the second period is the intersection of D_2 and S_2, and equilibrium in the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



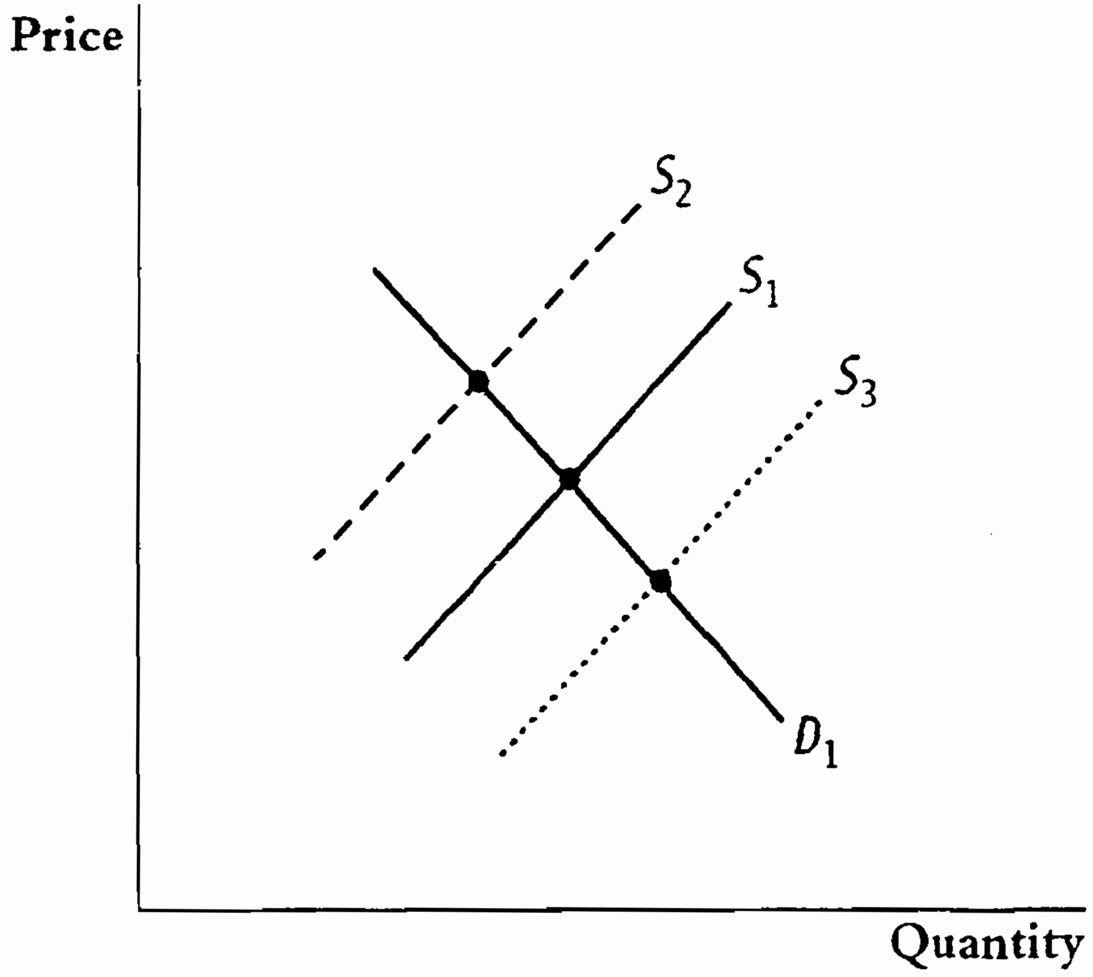
(b) Equilibrium price and quantity for 11 time periods

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 cigarette consumption data set

- The data set `cig_ch12.xlsx` 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.
`pop`	State population.
`packpc`	Number of packs per capita.
`income`	State personal income (total, nominal).
`tax`	Ave. state, federal, and ave. local excise taxes for fiscal year. (This is the cigarette-specific tax)
`avgprs`	Average price during fiscal year, including sales tax.
`taxs`	Average excise taxes for fiscal year, including sales tax. (This is the cigarette-specific tax + sales tax)

The TSLS estimation

- The original regression

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

- First stage of TSLS regression

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

- 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

```
open "@workdir/data/SW3/cig_ch12.xlsx"
setobs state year --panel-vars

genr salestax = (taxs - tax) / cpi
genr cigtax = tax / cpi
genr lnp = ln(avgprs / cpi)
genr lnq = ln(packpc)
genr lninc = ln(income / pop / cpi)

smpl year == 1995 --restrict --replace
```

Practice with the 1995 data

- Perform the OLS estimation of

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

```
ols lnq const lnp --robust
```

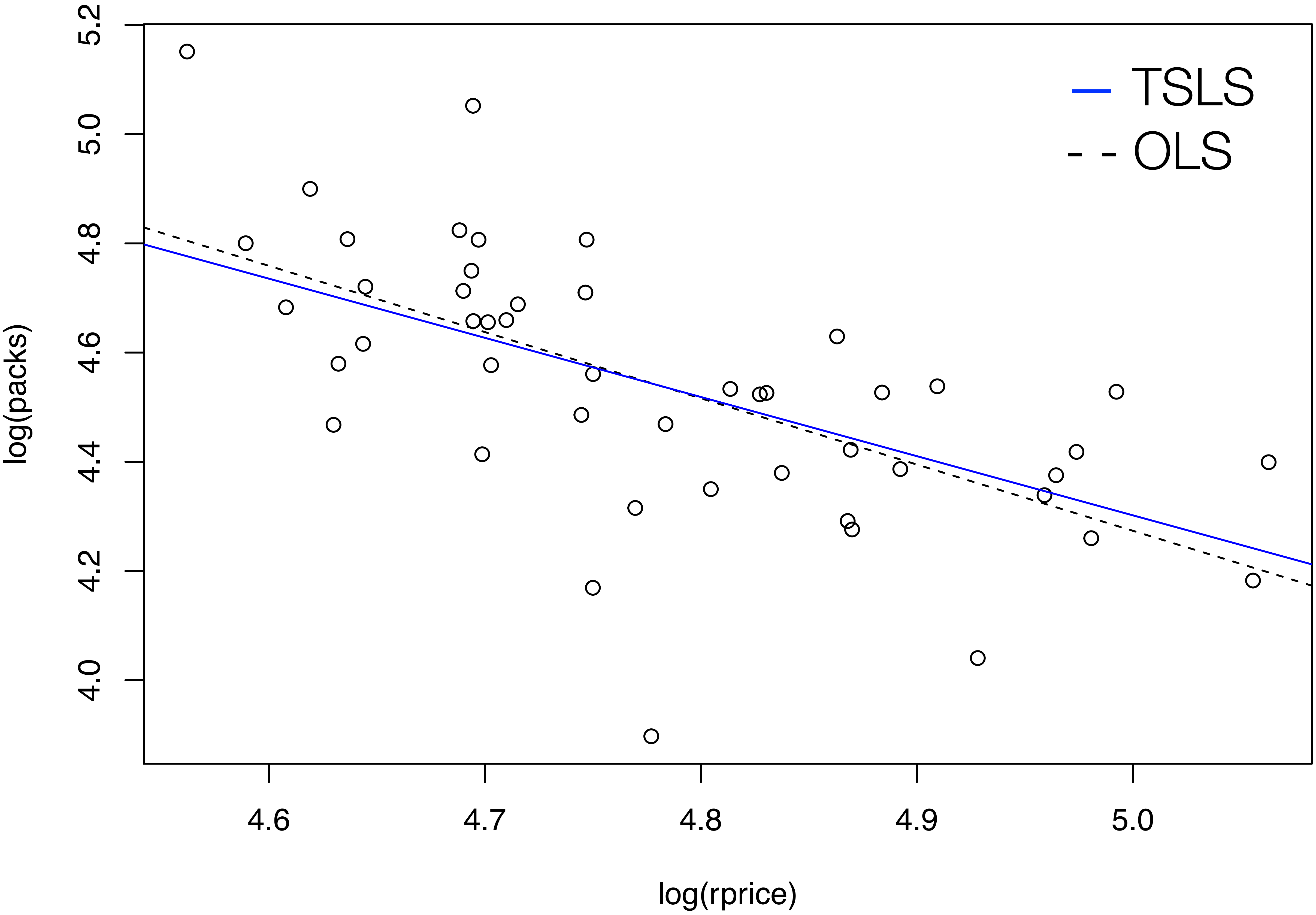
- Perform the TSLS estimation of the IV regression using OLS

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

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

```
genr lnphat = $yhat
ols lnq const lnphat --robust
```

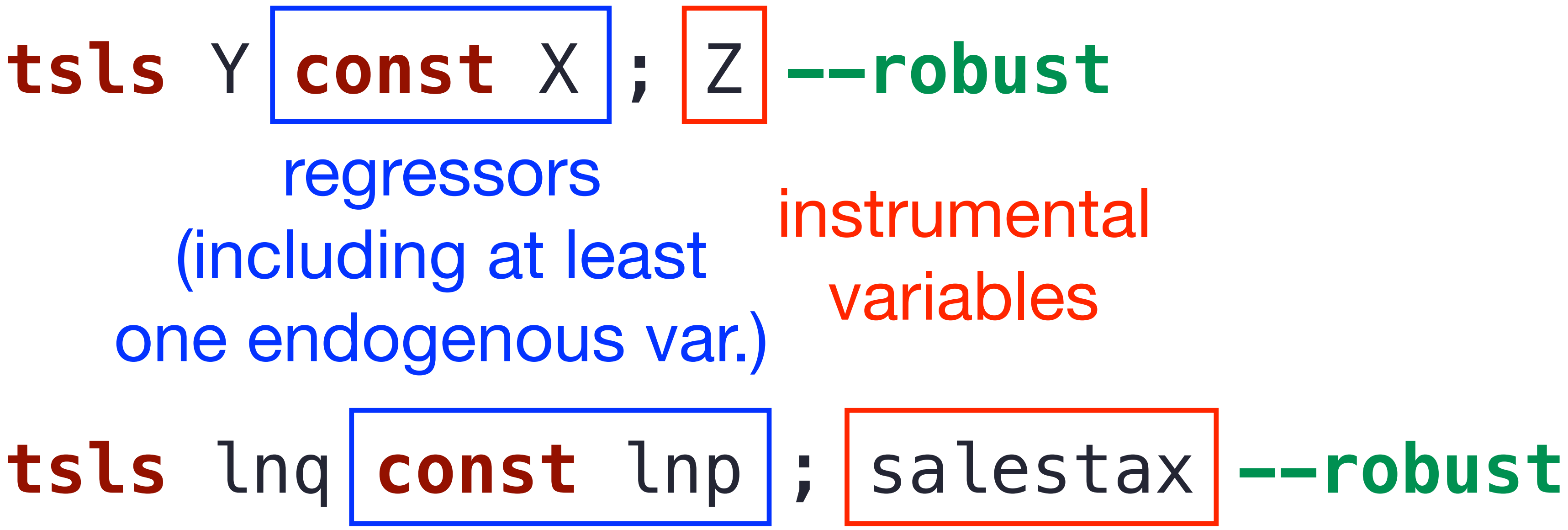
- Compare the two estimates $\hat{\beta}_1$ and $\hat{\beta}_1^{\text{TSLS}}$


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

The `tsls` command

- As we have seen, the TSLS method can be performed using the OLS method (`ols` command). However, the standard errors in the second stage is not correct.

- There is a single command `tsls`, which can provide both the correct TSLS estimates and standard errors.



```
tsls Y const X ; Z --robust
```

regressors (including at least one endogenous var.)

instrumental variables

```
tsls lnq const lnp ; salestax --robust
```

The second stage using OLS estimation

```
? ols lnq const lnphat --robust

Model 3: OLS, using observations 1-48
Dependent variable: lnq
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error     z      p-value
  -------------------------------------------------------
  const        9.71988      1.59712      6.086   1.16e-09 ***
  lnphat      -1.08359      0.333695    -3.247   0.0012   ***

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    2.358809   S.E. of regression   0.226447
R-squared            0.152490   Adjusted R-squared   0.134066
F(1, 46)             10.54455   P-value(F)           0.002178
Log-likelihood       4.204011   Akaike criterion    -4.408022
Schwarz criterion   -0.665620   Hannan-Quinn        -2.993763
```

The TSLS estimation using `tsls`

```
? tsls lnq const lnp ; salestax --robust

Model 4: TSLS, using observations 1-48
Dependent variable: lnq
Instrumented: lnp
Instruments: const salestax
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        9.71988      1.52832      6.360    8.35e-08 ***
  lnp         -1.08359      0.318918    -3.398    0.0014   ***

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    1.666792   S.E. of regression   0.190354
R-squared            0.405751   Adjusted R-squared   0.392832
F(1, 46)             11.54431   P-value(F)           0.001411
Log-likelihood      -34.38306   Akaike criterion     72.76611
Schwarz criterion    76.50851   Hannan-Quinn         74.18037
```

References

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