

计量经济学

第十一讲：工具变量回归（一）

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室 粤海校区汇文楼2613
E-mail huangjp@szu.edu.cn
Website <https://huangjp.com>

主要内容

- 内生性和外生性
 - 遗漏变量偏差
 - 双向因果关系
- 工具变量和 IV 估计量
 - 两阶段最小二乘估计
- 在香烟需求中的应用

内生性和外生性

内生性和外生性

Endogeneity and exogeneity

- 在回归模型中，当回归变量 X_k 与误差项 u 相关时， X_k 被称为**内生变量 (endogenous variable)**；若 X_k 与 u 不相关，则称其为**外生变量 (exogenous variable)**。
- 存在内生变量时，OLS 估计量是有偏的。
- 产生内生变量的主要原因：
 1. 存在遗漏变量（不可观测）；
 2. 存在测量误差（measurement error，详见第九章）；
 3. 存在双向因果关系（simultaneous causality）。

遗漏变量偏差（复习）

Omitted variable bias

- 如果回归变量与回归中漏掉的并对因变量起部分决定作用的某个变量（遗漏变量，omitted variable）相关，则 OLS 估计量有遗漏变量偏差（omitted variable bias）。

当满足下列条件时会产生遗漏变量偏差：

1. 遗漏变量和回归变量相关，
2. 遗漏变量是因变量的一个决定因素。

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

相关

其他变量

影响

例子：

1. 英语学习者百分比：满足 1 和 2；
2. 考试时间：不满足 1，但满足 2；
3. 每个学生的停车空间（教师停车场面积除以学生人数）：满足 1；但不满足 2。

遗漏变量偏差的影响（复习）

- 从第一条 OLS 假设 $E(u_i | X_i) = 0$ 可导出 $\text{corr}(X_i, u_i) = 0$ 。
- 如果存在遗漏变量偏差，则

$$\text{corr}(X_i, u_i) \neq 0 \quad \Rightarrow \quad E(u_i | X_i) \neq 0$$

- 若第一条 OLS 假设不成立，则 OLS 估计量是**有偏的**（因此称为遗漏变量偏差）。这个偏差即使在大样本下也不会消失，因此 OLS 估计量是**非一致的**。遗漏变量偏差下的 OLS 估计量满足

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}, \quad \text{此处 } \rho_{Xu} = \text{corr}(X_i, u_i)$$

如何应对遗漏变量偏差

- 遗漏变量是可观测的
 - 多元线性回归
- 遗漏变量是不可观测的，但不随时间变化（或不随个体变化）
 - 固定效应回归（面板数据）
- 遗漏变量是不可观测的，且随个体和时间变化，或者面板数据不可用
 - 本章介绍的工具变量回归

莫扎特效应

The “Mozart effect”

- 1993年发表在 *Nature* 上一项研究表明听莫扎特的音乐10-15分钟可以暂时提高你的 IQ 8-9分。

Rauscher, Shaw, and Ky, Music and spatial task performance, *Nature*, vol.365, pp.611, 1993.

- 莫扎特效应的依据是什么？
 - 多项研究显示，高中选修音乐或艺术课程的学生，其英语和数学成绩要高于没有选修这类课程的学生。
 - 但对这些研究的进一步深入调查发现测试成绩较好的真正原因与音乐或艺术课程没有关系。学习较好的学生有更多时间或更有兴趣选修这类课程，或者提供这类课程的学校都是较好的学校。
- 我们有理由认为存在遗漏变量偏差：学生天赋、学校质量等。

双向因果关系

Simultaneous causality

- 双向应归关系是指 X 决定 Y 的同时, Y 也决定 X 。
- 例如, 我们假定低学生教师比可以导致学生测试成绩的提升。但是, 如果政府对成绩较低的学校实施某种补助政策以增加教师数量, 那么低测试成绩也会导致低学生教师比。即学生教师比与学生测试成绩之间存在双向因果关系。
- 双向因果关系可以用联立方程的形式表示:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

由于 u_i 可以影响 Y_i , 而 Y_i 又影响 X_i , X_i 和 u_i 是相关的。

Philip Wright 的研究

- 最早提出工具变量估计的 Philip Wright 在20世纪20年代针对如何对动植物油（黄油、大豆等）设置进口关税进行研究。
- 理解关税带来的经济影响需要对对象商品的需求和供给曲线进行定量估计，具体地说就是需求和供给的价格弹性。
- 例如黄油的需求弹性问题可归结为下面的需求曲线

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



demand elasticity

需求弹性估计中存在的问题

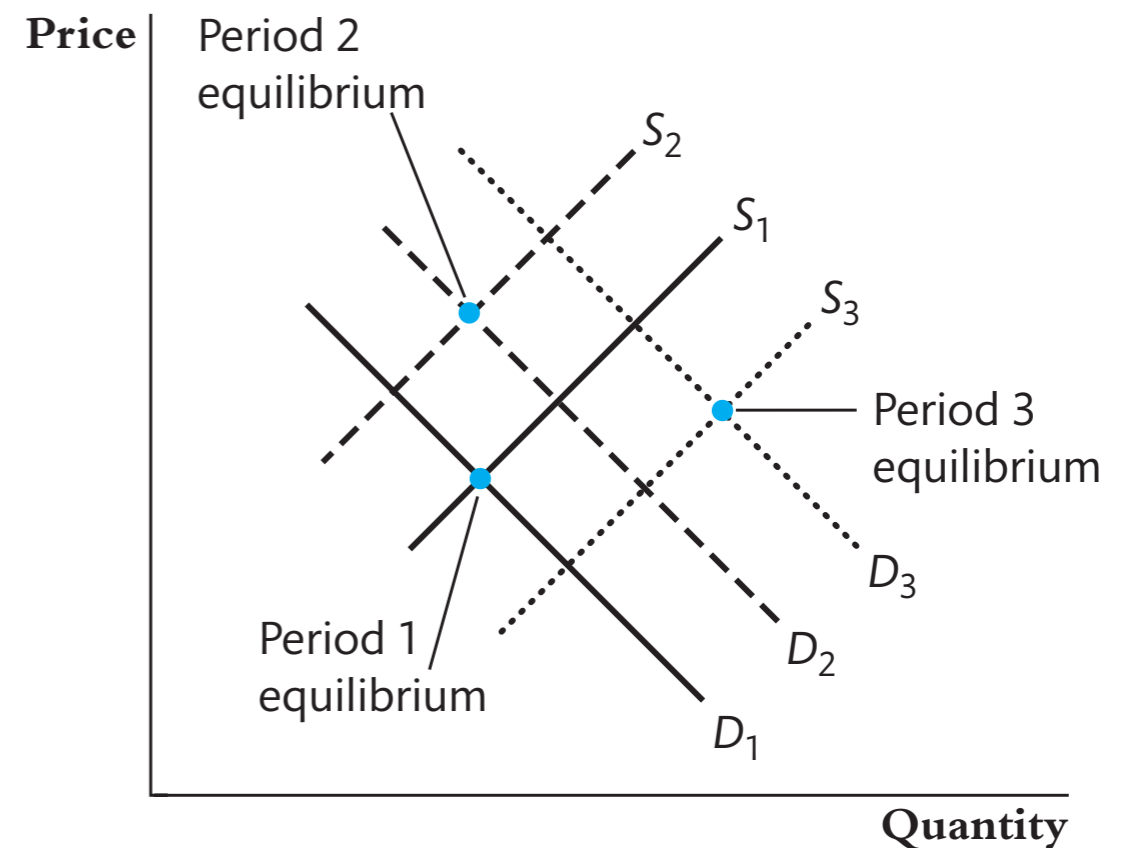
- Philip Wright 收集了美国 1912—1922年间黄油的年消费总量和年平均价格数据。用此数据可以很容易地用 OLS 估计出需求弹性。
- 然而，由于价格是需求和供给共同作用的结果，回归变量 $\ln(P_i^{\text{butter}})$ 可能与误差项相关。

需求与供给的相互作用

(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 .

第一期至第二期：需求增加、供给下降

第二期至第三期：需求增加、供给增加



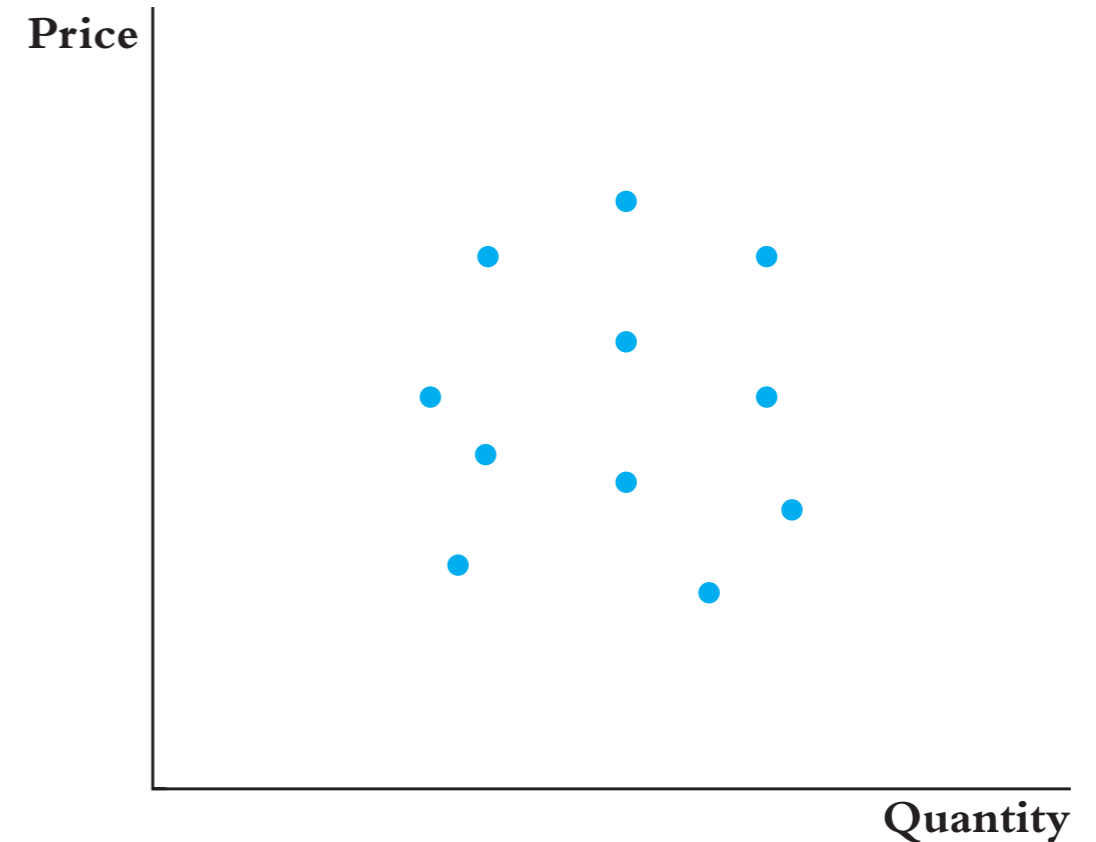
(a) Demand and supply in three time periods

需求与供给的相互作用

(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?

全11年的均衡价格与数量。

以此进行 OLS 拟合出的直线既不是供给曲线也不是需求曲线。



(b) Equilibrium price and quantity for 11 time periods

需求与供给的相互作用

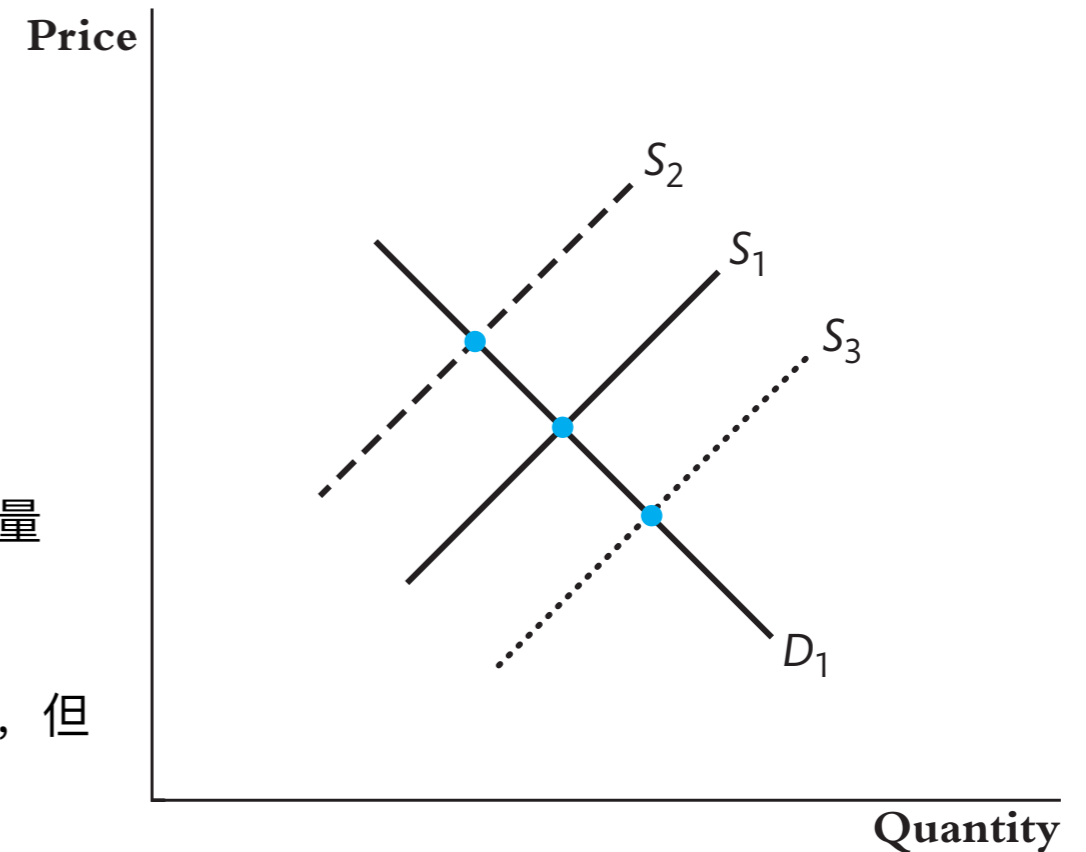
(c) When the supply curve shifts from S_1 to S_2 to S_3 but the demand curve remains at D_1 , the equilibrium prices and quantities trace out the demand curve.

如果能控制需求曲线不变，那么供给变化带来的均衡价格和均衡量的变化可以帮助我们描画需求曲线

一个办法是找到影响供给但不影响需求的第三个变量，即工具变量

Philip Wright 考虑的工具变量：天气

例如牧场降雨量的变化会影响牧草的生长，进而影响黄油的供给，但不会影响其需求。




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

工具变量和 IV 估计量

工具变量

Instrumental variable

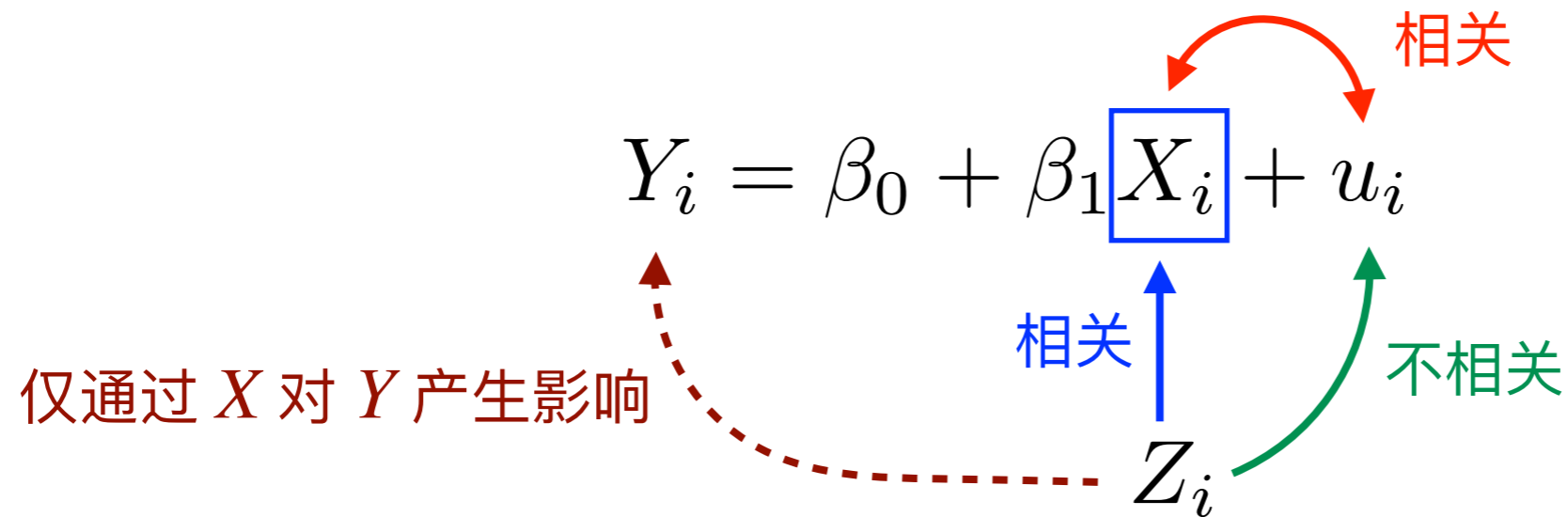
- 在一元线性回归模型中

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$


若 X_i 和 u_i 相关，则 OLS 估计量会发生偏差且不具备一致性。

- 此时可将 X_i 的变化分解成两部分：
 1. 和 u_i 相关的部分；
 2. 和 u_i 不相关的部分。
- 如果能找到一个**工具变量 (instrumental variable)** 并从中收集 X_i 变化中与 u_i 无关的部分，即将不产生偏差的部分分离出来，就能使回归系数的估计量具有一致性。

工具变量的有效条件



- 当 X 与 u 相关时，变量 Z 成为一个有效的工具变量的条件是
 1. **工具变量相关性 (instrument relevance)** : $\text{corr}(Z_i, X_i) \neq 0$
 2. **工具变量外生性 (instrument exogeneity)** : $\text{corr}(Z_i, u_i) = 0$

IV 估计量

IV estimator

- 在存在有效工具变量 Z_i 时,

$$\begin{aligned}\text{cov}(Z_i, Y_i) &= \text{cov}[Z_i, (\beta_0 + \beta_1 X_i + u_i)] \\ &= \beta_1 \text{cov}(Z_i, X_i) + \text{cov}(Z_i, u_i)\end{aligned}$$

由条件可知 $\text{cov}(Z_i, X_i) \neq 0$, $\text{cov}(Z_i, u_i) = 0$, 因此可导出

$$\beta_1 = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)}$$

- β_1 的 IV 估计量为

$$\hat{\beta}_1^{\text{OLS}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_1^{\text{IV}} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \beta_1$$

两阶段最小二乘估计量

The two stage least square (2SLS or TSLS) estimator

- 第一阶段 (first stage) : 用 Z_i 回归 X_i , 并计算预测值 \hat{X}_i

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

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

这里的 \hat{X}_i 就是 X_i 中随 Z_i 变化而变化的部分, 与 u_i 不相关。

- 第二阶段 (second stage) : 用 \hat{X}_i 回归 Y_i

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

由此得到的估计量 $\hat{\beta}_1^{\text{TSLS}}$ 就是 β_1 的 TSLS 估计量。

TSLS 估计量与 IV 估计量

- 根据 TSLS 的定义可以证明 (参见附录12.2)

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

- $\hat{\beta}_1^{\text{TSLS}}$ 具有一致性、在大样本下服从正态分布, 但有偏。

$$\hat{\beta}_1^{\text{TSLS}} \xrightarrow{p} \beta_1$$

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

在香烟需求中的应用

烟草税、香烟价格与香烟消费

- 众所周知，吸烟可能引起多种疾病，也会在不吸烟的人的生活造成不良影响（通过二手烟、社会成本的分担等）。
- 减少吸烟的一种方法是对香烟征收重税。但具体如何征收呢？例如，若要使香烟消费减少 20% 则香烟的税后售价应该是多少？
- 我们需要知道香烟的需求弹性。如果弹性为 -1 ，则价格上涨 20% 能使消费减少 20%。这时，烟草税应为售价的 20%。
- 和黄油一样，香烟的市场价格为内生变量，需要用到工具变量回归。

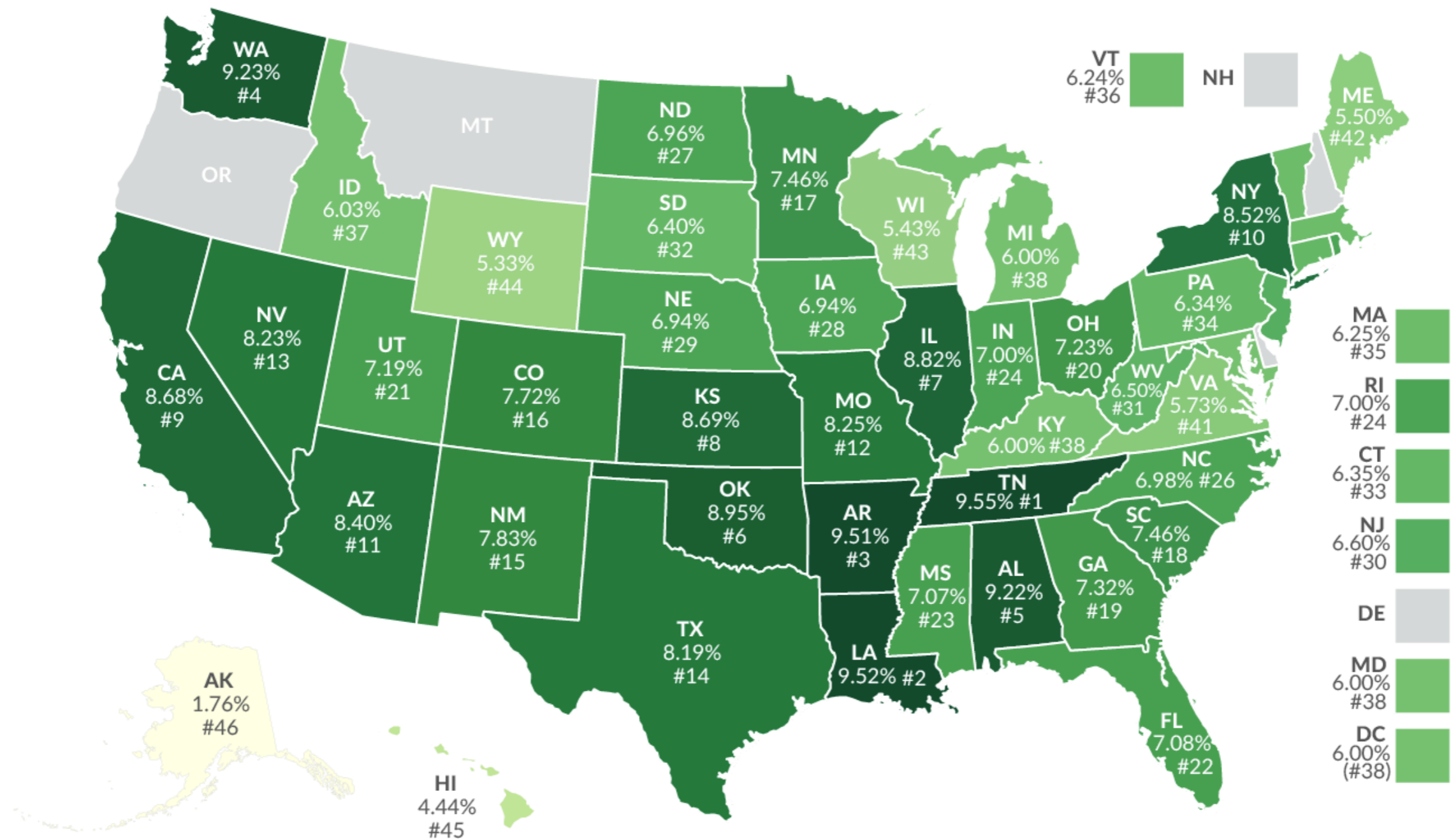
美国的税收制度

- 三种基本税种：基于收入，基于购买，基于产权
 - 基于收入：个人所得税、企业所得税、工薪税、资本利得税等
 - 基于购买：销售税（sales tax）、总收入税、增值税、特种消费税（excise tax）等
 - 基于产权：财产税、遗产税、财富税等
- 销售税（sales tax）：针对零售商品或服务，由州或地方政府征收。
- 特种消费税（excise tax）：针对特殊商品或服务（如烟草、酒精、汽油、赌博等），由联邦政府、州、地方政府征收。

参考：<https://taxfoundation.org/>

How High are Sales Taxes in Your State?

Combined State & Average Local Sales Tax Rates, January 2021



Notes: City, county and municipal rates vary. These rates are weighted by population to compute an average local tax rate. The sales taxes in Hawaii, New Mexico, and South Dakota have broad bases that include many business-to-business services. D.C.'s rank does not affect states' ranks, but the figure in parentheses indicates where it would rank if included.

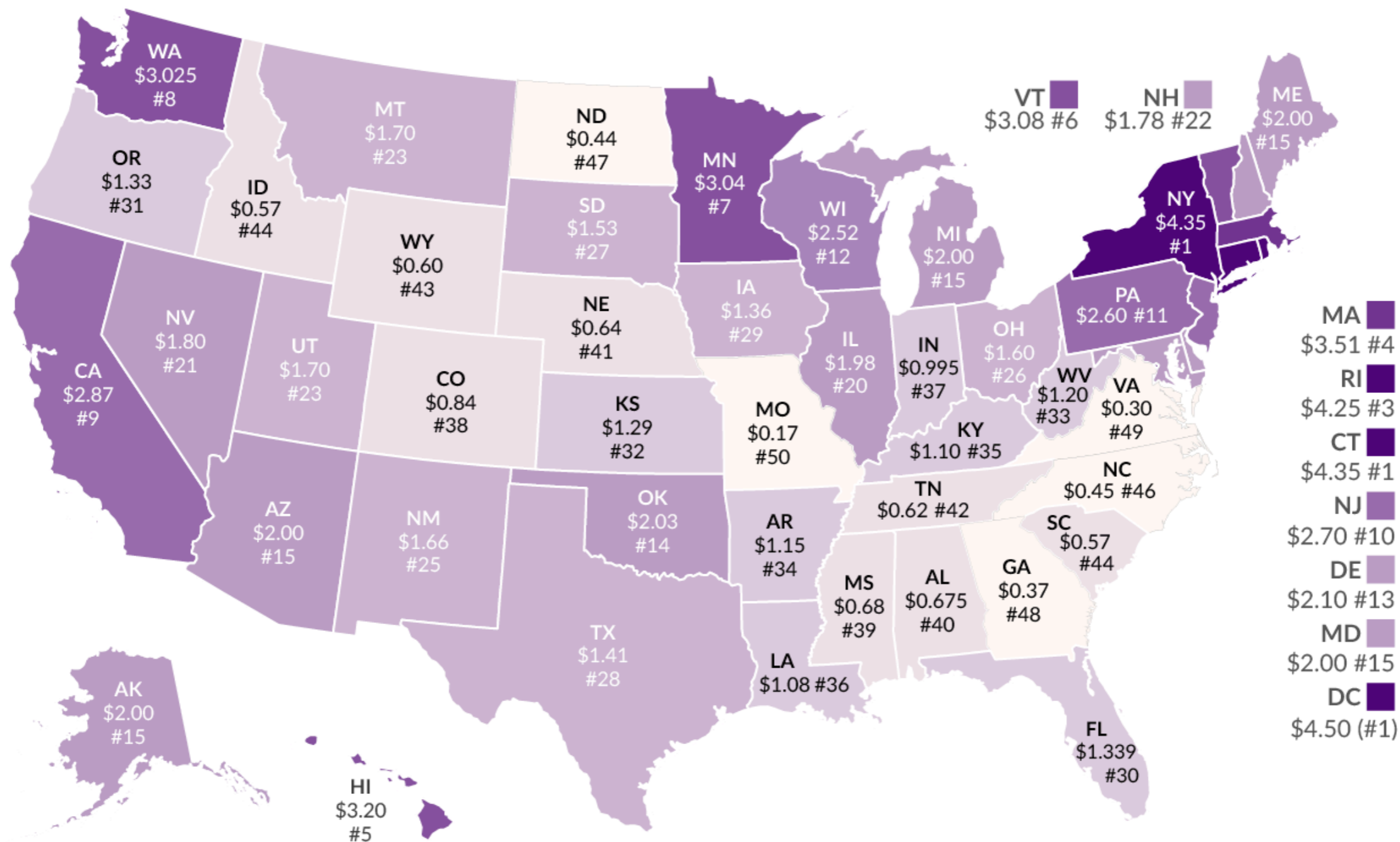
Sources: Sales Tax Clearinghouse; Tax Foundation calculations; State Revenue Department websites

Combined State & Average Local Sales Tax Rates



How High Are Cigarette Taxes in Your State?

State Cigarette Excise Tax Rates (dollars per 20-pack), January 2019



Note: Local taxes are not included and can be substantial. D.C.'s rank does not affect states' ranks, but the figure in parentheses indicates where it would rank if included.

Source: Bloomberg Tax; state statutes.

State Cigarette Excise Tax Rates (dollars per 20-pack)



香烟消费数据集

- 数据文件 `cig_ch12.xlsx` 中包含美国 48 个大陆州 1985 和 1995 年的香烟消费数据。

- 除了 `state` 和 `year`, 该数据集还包含其他 7 个变量:

<code>cpi</code>	Consumer price index.
<code>pop</code>	State population.
<code>packpc</code>	Number of packs per capita.
<code>income</code>	State personal income (total, nominal).
<code>tax</code>	Ave. state, federal, and ave. local excise taxes for fiscal year. (This is the cigarette-specific tax)
<code>avgprs</code>	Average price during fiscal year, including sales tax.
<code>taxs</code>	Average excise taxes for fiscal year, including sales tax. (This is the cigarette-specific tax + sales tax)

工具变量回归设定

- 因变量：州内人均香烟消费量（包/人） $Q_i^{\text{cigarettes}}$
- 回归变量：每包香烟的平均销售价格（含烟草税和销售税） $P_i^{\text{cigarettes}}$
- 工具变量：销售税 SalesTax_i
- 工具变量是否有效？
 - 工具变量相关性：销售税直接影响销售价格
 - 工具变量外生性：各州设定销售税率不同，其原因主要是各州在决定不同税种占比时的考量不同，其中关于公共财政的决策受到政治考量的驱使，而基本不受香烟需求的影响。（也不排除香烟生产企业对政治家施加压力而影响立法的可能性）

TSLS 估计

- 原始回归模型

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

No. of packs per capita
in the state

ave. price per pack

- 第一阶段

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

calculated as (taxes - tax)

- 第二阶段

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

TSLS 估计

基于 1995 年的横截面数据

- 数据整理

```
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
```

TOLS 估计

基于 1995 年的横截面数据

- OLS 回归

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

ols lnq **const** lnp **--robust**

- IV 回归的 TOLS 估计 (两阶段 OLS)

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

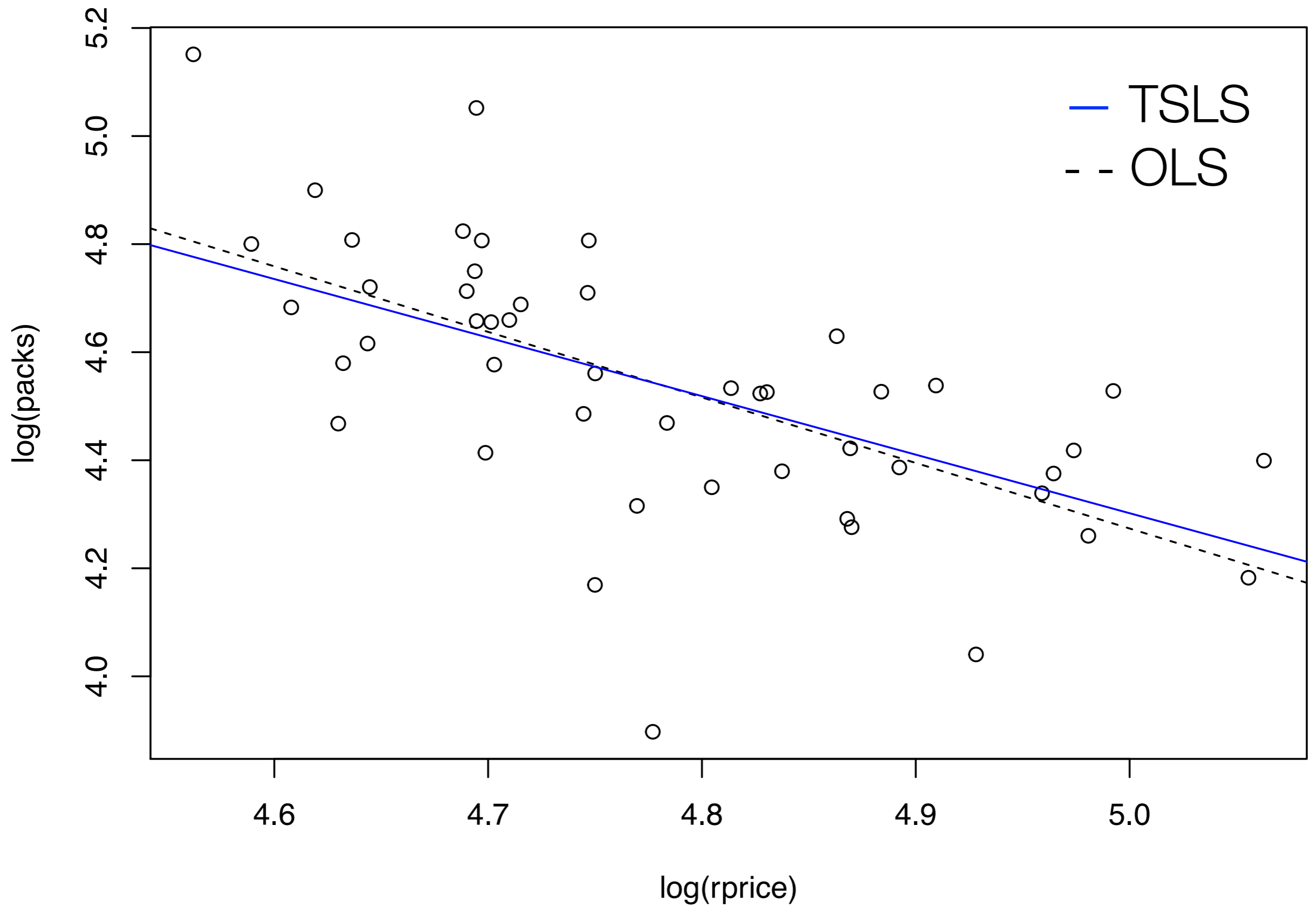
$$\ln(Q_i^{\text{cigarettes}}) = \beta_0^{\text{TOLS}} + \beta_1^{\text{TOLS}} \ln(\widehat{P}_i^{\text{cigarettes}}) + u_i^{\text{TOLS}}$$

ols lnp **const** salestax **--robust**

genr lnphat = **\$yhat**

ols lnq **const** lnphat **--robust**

- 比较 $\hat{\beta}_1^{\text{OSL}}$ 和 $\hat{\beta}_1^{\text{TOLS}}$



tsls 命令

- TSLS 估计可以通过两次 OLS 估计（用 `ols` 命令）实现。但是，第二阶段估计的标准误不正确。这是因为计量软件不知道该回归是第二阶段。
- 在 `gretl` 中针对 TSLS 估计的专用命令是 `tsls`。该命令可以正确计算标准误。

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

回归变量
(至少包含一个内生变量)

以 ; 分隔
工具变量

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

用 `ols` 命令估计第二阶段的结果

```
? 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

用 `tsls` 命令的估计结果

```
? 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