计量经济学

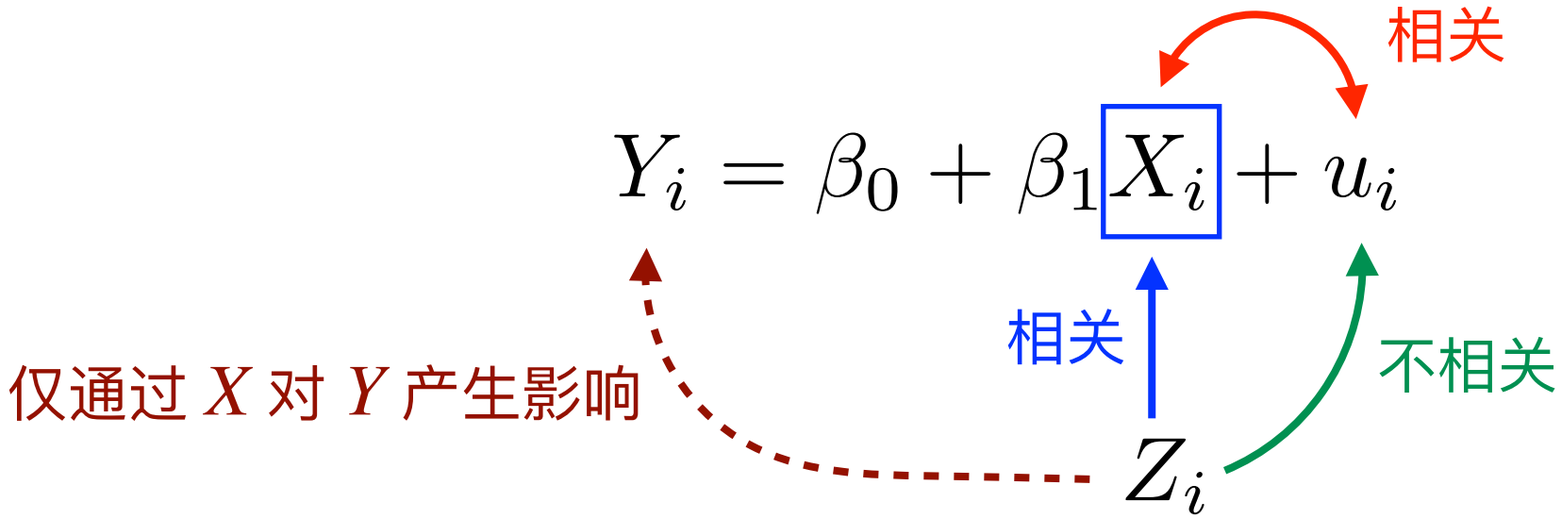
第十一讲：工具变量回归（二）

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

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,

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

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

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

- β_1 的 IV 估计量为

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

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

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。

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：

`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)

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。

以；分隔

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

回归变量（至少包含一个内生变量）　　工具变量

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

主要内容

- 一般 IV 回归模型
  - 工具变量的有效性检验
  - 回归变量的内生性检验
- 在香烟需求中的应用（面板数据）
- 有效工具变量的来源

一般 IV 回归模型

一般工具变量回归模型

- 一般工具变量回归模型包含四种变量：

  因变量 Y，内生变量 X，外生变量 W，工具变量 Z。

- W 与 Z 都是外生变量（与 u 不相关），其区别是 W 可以直接影响 Y，因此被包含在模型中；而 Z 只能通过 X 间接影响 Y。同理，W 不能做为工具变量。

- 在一般情况下，X、W 和 Z 都可以取复数个。但 Z 的个数不能小于 X 的个数。

一般工具变量回归模型

- 一般工具变量回归模型：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots \beta_{k+r} W_{ri} + u_i$$

  $Y_i, X_{\bullet i}, W_{\bullet i}$ 分别为因变量、内生变量和外生变量。$Z_{1i}, \ldots, Z_{mi}$ 为 m 个工具变量。

- 若 $m > k$，则称回归系数是**过度识别的（over-identified）**；若 $m = k$，则称回归系数是**恰好识别的（exactly identified）**；若 $m < k$，则称回归系数是**不可识别的（under-identified）**。

- IV 模型的估计必须是恰好识别或过度识别的。

一般 IV 模型的 TSLS 估计

- 结构方程（structural equation）

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots \beta_{k+r} W_{ri} + u_i$$

- 第一阶段回归（first stage regression）

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

tsls 命令

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

结构方程中包含的外生变量
应出现在；的两边

练习

- 将人均收入的对数做为外生变量加入香烟需求弹性的回归模型中，复制书中的 (12.15)

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

- 将销售税和烟草税做为工具变量，复制书中的 (12.16)

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

工具变量有效的两个条件

- 工具变量相关性

  给定外生变量 W 时，工具变量必须能够对内生变量的变动有足够的解释能力，且在第二阶段回归中不存在完全多重共线性。

  如果工具变量对内生变量的解释能力不足，则称其为**弱工具变量（weak instruments）**。弱工具变量会使 TSLS 估计量有偏。

- 工具变量外生性

  工具变量与误差项不相关，即 $\text{corr}(Z_{ji}, u_i) \neq 0,\ j = 1, \ldots, m$

弱工具变量的检验

- 当只有一个内生变量时

  可以用第一阶段回归中检验 “所有工具变量系数都为零” 的 F 统计量做为参考。常用的经验法则是当 F 统计量取值小于 10 时，表明工具变量是弱的。

- 存在弱工具变量时的对策

  在过度识别的情况下，可以选择去除较弱的工具变量。
  如果无法去除弱工具变量（恰好识别、或过度识别时强工具变量不够用），则应考虑：1. 选择其他更好的工具变量；2. 保留弱工具变量但选择 TSLS 以外对弱工具变量不敏感的估计方法（附录 12.5）。

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

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

            coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        9.43066      1.25939      7.488    1.93e-09 ***
  lnp         -1.14338      0.372303    -3.071    0.0036   ***
  lninc        0.214515     0.311747     0.6881   0.4949

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    1.617235   S.E. of regression   0.189575
R-squared            0.430985   Adjusted R-squared   0.405696
F(2, 45)             8.191141   P-value(F)           0.000925
Log-likelihood      -23.67640   Akaike criterion     53.35280
Schwarz criterion    58.96640   Hannan-Quinn         55.47419

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 1.20218
  with p-value = 0.272886

Weak instrument test -
  First-stage F-statistic (1, 45) = 44.7305
  A value < 10 may indicate weak instruments
```

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

```
Model 2: TSLS, using observations 1-48
Dependent variable: lnq
Instrumented: lnp
Instruments: const lninc salestax cigtax
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        9.89496      0.959217    10.32     1.95e-13 ***
  lnp         -1.27742      0.249610    -5.118    6.21e-06 ***
  lninc        0.280405     0.253890     1.104    0.2753

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    1.588044   S.E. of regression   0.187856
R-squared            0.432398   Adjusted R-squared   0.407171
F(2, 45)             16.17491   P-value(F)           5.09e-06

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 3.34671
  with p-value = 0.0673395

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 0.332622
  with p-value = P(Chi-square(1) > 0.332622) = 0.564119

Weak instrument test -
  First-stage F-statistic (2, 44) = 209.676
  A value < 10 may indicate weak instruments
```

工具变量外生性的检验

- 弱工具变量会使 TSLS 估计量有偏，而工具变量外生性会使 IV 回归无法得到一致的估计量。
- 我们能从统计上检验工具变量外生性的假设吗?
- 在恰好识别时**无法检验**。此时只能凭专业知识和对问题的理解进行判断。
- 在过度识别时，可以用**过度识别约束检验（test of over-identifiying restrictions，J 统计量）**检验是否所有的工具变量都是外生的。

  具体方法是，令 $\hat{u}_i^{\text{TSLS}}$ 为 IV 回归的 TSLS 估计残差，并用工具变量 Z 和 外生变量 W 对其进行 OLS 回归。令 F 表示该回归中工具变量系数都为零的联合假设的同方差适用 F 统计量，则过度识别约束检验统计量为 $J = mF$，J 在大样本下服从 χ^2_{m-k} 分布。

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

```
Model 2: TSLS, using observations 1-48
Dependent variable: lnq
Instrumented: lnp
Instruments: const lninc salestax cigtax
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        9.89496      0.959217    10.32     1.95e-13 ***
  lnp         -1.27742      0.249610    -5.118    6.21e-06 ***
  lninc        0.280405     0.253890     1.104    0.2753

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    1.588044   S.E. of regression   0.187856
R-squared            0.432398   Adjusted R-squared   0.407171
F(2, 45)             16.17491   P-value(F)           5.09e-06

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 3.34671
  with p-value = 0.0673395

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 0.332622
  with p-value = P(Chi-square(1) > 0.332622) = 0.564119

Weak instrument test -
  First-stage F-statistic (2, 44) = 209.676
  A value < 10 may indicate weak instruments
```

This is another test

在 gretl 中计算 J 统计量和 p 值

- 在运行 TSLS 估计之后，执行下列命令

```
genr esterr = $uhat
ols esterr const lninc salestax cigtax
restrict
  b[3] = 0
  b[4] = 0
end restrict
scalar Jstat = 2 * $test
scalar Jpvalue = pvalue(X, 1, Jstat)
print Jstat Jpvalue
```

- 其结果为

```
Jstat =  0.30703124
Jpvalue =  0.57950767
```

IV 模型中内生性的检验

- 当 IV 回归模型中的内生变量 X 实际上是外生的时候，TSLS 估计量相对于 OLS 估计量效率更低（标准误更大）。

- 通过对回归模型（结构方程）分别进行 OLS 回归和 TSLS 回归，并比较内生变量系数的估值结果，可以判断内生变量是否真的是内生的。
  - 如果两种估计结果差别很大，说明 TSLS 有效，即 X 为内生。
  - 如果两种估计结果相似，说明所有变量都是外生。

- 这种检验被称为 **Durbin-Wu-Hausman 检验（DWH 检验，或者 Hausman 检验）**。

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

```
Model 2: TSLS, using observations 1-48
Dependent variable: lnq
Instrumented: lnp
Instruments: const lninc salestax cigtax
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        9.89496      0.959217    10.32     1.95e-13 ***
  lnp         -1.27742      0.249610    -5.118    6.21e-06 ***
  lninc        0.280405     0.253890     1.104    0.2753

Mean dependent var   4.538837   S.D. dependent var   0.243346
Sum squared resid    1.588044   S.E. of regression   0.187856
R-squared            0.432398   Adjusted R-squared   0.407171
F(2, 45)             16.17491   P-value(F)           5.09e-06

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 3.34671
  with p-value = 0.0673395

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 0.332622
  with p-value = P(Chi-square(1) > 0.332622) = 0.564119

Weak instrument test -
  First-stage F-statistic (2, 44) = 209.676
  A value < 10 may indicate weak instruments
```

在香烟需求中的应用（面板数据）

香烟需求的面板数据分析

- 在香烟需求方程中，有可能存在其他的遗漏变量。
- 除了价格以为还有什么因素能影响香烟消费呢？一个是收入，我们已经将其加入到回归模型中。另一个是需求的历史因素，例如当地是否有种植烟草的历史，而如果烟草种植业是当地的重要产业，则当地的烟草税就可能更低。
- 因为我们有面板数据，因此可以将个体固定效应移除（如各州的历史、自然、文化特征）。
- 如果时间跨度较大，那么估计出的需求弹性是长期需求弹性。因为吸烟会使人上瘾，当价格变化时需求不会马上随着改变，因此短期内的需求弹性会小于长期的需求弹性。

香烟需求的面板数据分析

基于 1985 和 1995 年的数据

- 利用差分的结构方程

$$\ln(Q^{\text{cig}}_{i,1995}) - \ln(Q^{\text{cig}}_{i,1985}) = \beta_0 + \beta_1 \Big[\ln(P^{\text{cig}}_{i,1995}) - \ln(P^{\text{cig}}_{i,1985}) \Big] + \beta_2 \Big[\ln(\text{Income}_{i,1995}) - \ln(\text{Income}_{i,1985}) \Big] + u_i$$

其中价格的对数差分项为内生变量，收入的对数差分项为外生变量。工具变量为

$$\text{SalesTax}_{i,1995} - \text{SalesTax}_{i,1985}$$

$$\text{CigTax}_{i,1995} - \text{CigTax}_{i,1985}$$

TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q^{cigarettes}_{i,1995}) - \ln(Q^{cigarettes}_{i,1985})$

Regressor	(1)	(2)	(3)
$\ln(P^{cigarettes}_{i,1995}) - \ln(P^{cigarettes}_{i,1985})$	−0.94 (0.21) [−1.36, −0.52]	−1.34 (0.23) [−1.80, −0.88]	−1.20 (0.20) [−1.60, −0.81]
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34) [−0.16, 1.21]	0.43 (0.30) [−0.16, 1.02]	0.46 (0.31) [−0.16, 1.09]
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage *F*-statistic	33.7	107.2	88.6
Overidentifying restrictions *J*-test and *p*-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The *J*-test of overidentifying restrictions is described in Key Concept 12.6 (its *p*-value is given in parentheses), and the first-stage *F*-statistic is described in Key Concept 12.5. Heteroskedasticity-robust standard errors are given in parentheses beneath coefficients, and 95% confidence intervals are given in brackets.

J 检验拒绝了工具变量全部有效（外生）的原假设。但是我们无法通过统计结果判断哪个工具变量有效、或者两个都无效。
一般情况下，销售税为外生的可能性大于烟草专项税。

练习：复制表 12.1 的回归结果

- 差分可以用 `diff` 命令获得
- 在运行 `diff` 命令后将样本限定为 1995 年。这样可以获得和书中一致的标准误。
- J 统计量和 p 值按照前述方法计算。

有效工具变量的来源

如何选择合适的工具变量

- 工具变量回归理论上简单且功能强大，但在实践中，最困难的部分就是如何找到有效的（既相关又外生）工具变量。
- 主要方法有两种：
  - 利用经济理论
  - 利用专业知识、对问题的深入了解、以及对数据细节的关注
- 建议：广泛阅读优质文献（包括书和论文），积极和导师、同学、同行进行讨论。

实例（一）

- 把罪犯关进监狱会减少犯罪吗?

  因变量：犯罪率
  回归变量：监禁率
  控制变量：经济环境变量、人口统计变量、等

- **存在双向因果偏差**：

  如果犯罪率上升，会有更多警力投入到执法活动中，导致监禁率上升

- 工具变量：监狱容量
  → 针对减少监狱过分拥挤的诉讼（Levitt, 1996）

实例（二）

- 缩小班级规模能提高测试成绩吗?

  因变量：测试成绩
  回归变量：班级大小
  控制变量：学生家庭收入、英语能力、等

- **存在不可观测的遗漏变量**：

  校外学习机会、父母对学习的兴趣、教师能力等

- 工具变量：学生生日
  → 出生时间是随机的，因此学区内幼儿园的潜在入学人数可以做为工具变量的候选
  (Hoxby, 2000)

实例（三）

- 对心脏病的积极治疗能延长寿命吗?

  因变量：患者的寿命
  回归变量：患者是否接受心导管术（二值变量）
  控制变量：患者年龄、体重、其他健康状况指标等

- **存在选择偏差**：

  是否接受治疗的决定不是随机的（和遗漏的健康状态变量相关）

- 工具变量：患者居住的地理位置
  → 患者距最近的心导管术医院的相对距离
  （McClellan, McNeil, and Newhouse, 1994）