Econometrics 1

Lecture 13: Instrumental Variables (2)

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`

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

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

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

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

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

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` 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 general IV regression model

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:

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

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.

Identification

- Exact identification
  The number of instruments (m) equals the number of endogenous regressors (k): $m = k$.

- Over-identification
  The number of instruments (m) exceeds the number of endogenous regressors (k): $m > k$.

- Under-identification
  The number of instruments (m) is less than the number of endogenous regressors (k): $m < k$.

TSLS in the general IV model

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

- First stage estimation

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

The `tsls` command with included exogenous variables

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

included exogenous variables
on both sides of " **;** "

Practice

- Take the logarithm of real income per capita as an included exogenous variable (i.e., W) and reproduce Equation (12.15).

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

- Take the sales tax and the cigarettes specified tax as two IVs and reproduce Equation (12.16).

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

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 — instrument relevance

- 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, and in addition there is no perfect multicollinearity in the second stage.

  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.

Checking for weak instruments

- Weak instruments — a rule of thumb

  When there is a single endogenous regressor, 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.

```
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
```

The validity of IV — instrument exogeneity

- **Instrument exogeneity** — the instruments are uncorrelated with the error term.

  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.

Test of overidentifying restrictions

- Let $\hat{u}_i^{\mathrm{TSLS}}$ be the residuals from TSLS estimation of

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

- Use OLS to estimate the regression coefficients in

$$\hat{u}_i^{\mathrm{TSLS}} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} \\ + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i$$

- Let F denote the homoskedasticity-only *F*-statistic testing $\delta_1 = \cdots = \delta_m = 0$. The overidentifying restriction test statistic is $J = mF$.

- In large samples J is distributed χ^2_{m-k}.

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

Evaluate the *J*-statistic and the corresponding *p*-value

- After running the TSLS estimation, run the following

```
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
```

- This produces

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

Application to the demand for cigarette with panel data

- There might be other omitted variables in the cigarette demand regression, such as if a state grows tobacco.
- Such state fixed effects can be removed using panel data.
- Using the differences of data in 1985 and 1995, we can remove state fixed effects and investigate the long run elasticity (since smoking is addictive).

Application to the demand for cigarette with panel data

- The model using 10-year changes

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

where the difference of log price is the endogenous, the difference of log income is exogenous, with instruments being

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

and/or

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

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

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

Regressor	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})$	−0.94** (0.21)	−1.34** (0.23)	−1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
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	33.70	107.20	88.60
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. Individual coefficients are statistically significant at the *5% level or **1% significance level.

Practice: reproduce Table 12.1

- The differences can be obtained using `diff` command.

- After taking the differences, restrict the sample to `year` = 1995. This makes the robust standard errors coincide with those in the book.

- Check the *J*-statistic using the method given in the book.

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 books and papers, discuss with supervisors or colleagues.

Examples of IV choice

- Does putting criminals in jail reduce crime?

  Dependent Var. = crime rates
  Independent Var. = incarceration rates
  (fractions of prisoners in the population)
  Control Var. = economic conditions,
  demographics,
  etc.

- **Simultaneous causality bias**

- Instruments: capacity of existing prisons
  → lawsuits aimed at reducing prison overcrowding
  (Levitt, 1996)

Examples of IV choice

- Does cutting class sizes increase test scores?

  Dependent Var. = test scores
  Independent Var. = class sizes
  Control Var. = student affluence,
  ability to speak English,
  etc.

- **Unavailable omitted variables**:
  outside learning opportunities, parental interest in learning, quality of teachers, etc.

- Instruments: timing of birth (Hoxby, 2000)

Examples of IV choice

- Does aggressive treatment of heart attacks prolong lives?

  Dependent Var. = length of survival of patient
  Independent Var. = whether the patient received
  cardiac catheterization* (binary treatment)
  Control Var. = age, weight,
  other measured health conditions,
  etc.

- **Selection bias**: the treatment decision is not random
  (related to omitted health factors)

- Instruments: difference between the distance from patient home to the nearest cardiac catheterization hospital and the distance to the nearest other general hospital (0 if CC hospital is nearest, and positive otherwise) (McClellan, McNeil, & Newhouse, 1994)

* Cardiac catheterization is a procedure in which a catheter, or tube, is inserted into a blood vessel and guided all the way to the heart to obtain information about the heart and coronary arteries.

Summary

- IV: a variable that is correlated with the endogenous var. but uncorrelated with the error term.
- TSLS is a useful tool.
- Weak instruments:
  TSLS estimator can be biased even in large sample.
  Can be checked using first stage *F* if there is only one endogenous var.
- Not exogenous instruments:
  TSLS estimator is inconsistent.
  Can be checked using *J*-stat for over-identification.

References

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