Lecture 10: Instrumental Variables

Introduction ot Econometrics,Fall 2018

Zhaopeng Qu

Nanjing University

11/22/2018

1. Review Previous Lecture of Internal Validity

2. Instrumental Variable Method

3. Checking Instrument Validity

4. Instrumental Variable for multiple regression

5. Review the last lecture

6. IV with Heterogeneous Causal Effects: Simple Case

7. IV with Heterogeneous Causal Effects: Generalization

8. Some Practical Guide by Angrist and Pischke (2010)

Review Previous Lecture of Internal Validity

Threatens to Internal Validity

- Three endogenous in OLS regression are:
  - **Omitted Variable Bias**(a variable that is correlated with X but is unobserved)
  - **Simultaneity or reverse causality Bias** (X causes Y,Y causes X)
  - **Errors-in-Variables Bias** (X is measured with error)
- One easy way to deal with these endogouneity is using instrumental variable.

Instrumental Variable Method

Introduction

- The earliest application involved attempts to estimate demand and supply curve for product.
- A simple but difficult question: How to find the supply or demand curves?
- Difficulty: We can only observe intersections of supply and demand, yielding pairs.
- Solution: Wright(1928) use variables that appear in one equation to shift this equation and trace out the other.
- The variables that do the shifting came to be known as **Instrumental Variables** method.
- It is well-known that IV can address the problems of omitted variable bias, measurement error and reverse causality problems.

Terminology: endogeneity and exogeneity

- An *endogenous variable* is one that both we are interested in and is correlated with u.
- An *exogenous variable* is one that is uncorrelated with u.
- Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality.
- However, this definition is narrow and IV regression can be used to address OVB and errors-in-variable bias, not just to simultaneous causality bias.

Instrumental variables: 1 endogenous regressor & 1 instrument

- suppose a simple OLS regression like previous equation

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

- Because $E[u_i|X_i] \neq 0$, then we can use an instrumental variable(Z_i) to obtain an consistent estimate of coefficient.
- Intuitively, we want to split X_i into two parts:
  1. part that is correlated with the error term.
  2. part that is uncorrelated with the error term.
- If we can isolate the variation in X_i that is uncorrelated with u_i,then we can use this part to obtain a consistent estimate of the causal effect of X_i on Y_i.

Instrumental variables: 1 endogenous regressor & 1 instrument

- An instrumental variable Z_i must satisfy the following 2 properties:

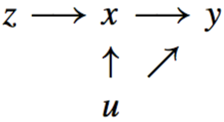
  1. **Instrumental relevance**: Z_i should be **correlated** with the casual variable of interest, X_i (endogenous variable),thus

     $$Cov(X_i, Z_i) \neq 0$$

     .

  2. **Instumental exogeneity**: Z_i is as good as randomly assigned and Z_i only affect on Y_i through X_i affecting Y_i channel.

     $$Cov(Z_i, u_i) = 0$$



IV estimator: Jargon

- Our simple OLS regression: Causal relationship of interest

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

- **First-Stage** regression: regress *endogenous variable* on IV

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

- **Reduced-Form**: regress outcome variable on IV

$$Y_i = \delta_0 + \delta_1 Z_i + e_i$$

IV estimator: Two Steps Least Square (2SLS)

- We can estimate the causal effect of X_i on Y_i in two steps

1. **First stage**: Regress X_i on Z_i & obtain predicted values of $\hat{X}_i$,if $Cov(Z_i, u_i) = 0$, then $\hat{X}_i$ contains variation in X_i that is uncorrelated with u_i

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

.

2. **Second stage**: Regress Y_i on $\hat{X}_i$ to obtain the Two Stage Least Squares estimator $\hat{\beta}_{2SLS}$

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{\hat{X}})}{\sum(\hat{X}_i - \bar{\hat{X}})^2}$$

IV estimator: Two Steps Least Square (2SLS)

- we substitute

$$\hat{X}_i - \overline{\hat{X}} = \hat{\pi}_1(Z_i - \bar{Z})$$

- then we could obtain

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum(\hat{X}_i - \overline{\hat{X}})^2}$$

IV estimator: Two Steps Least Square (2SLS)

- we substitute

$$\hat{X}_i - \overline{\hat{X}} = \hat{\pi}_1(Z_i - \bar{Z})$$

- then we could obtain

$$\begin{aligned}\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum(\hat{X}_i - \overline{\hat{X}})^2} \\ &= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \overline{Z})}{\sum \hat{\pi}_1^2(Z_i - \overline{Z})^2}\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- we substitute

$$\hat{X}_i - \overline{\hat{X}} = \hat{\pi}_1(Z_i - \bar{Z})$$

- then we could obtain

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum(\hat{X}_i - \overline{\hat{X}})^2} \\
&= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \overline{Z})}{\sum \hat{\pi}_1^2(Z_i - \overline{Z})^2} \\
&= \frac{1}{\hat{\pi}_1}\frac{\sum(Y_i - \bar{Y})(Z_i - \overline{Z})}{\sum(Z_i - \overline{Z})^2}
\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- we substitute

$$\hat{X}_i - \overline{\hat{X}} = \hat{\pi}_1(Z_i - \bar{Z})$$

- then we could obtain

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum(\hat{X}_i - \overline{\hat{X}})^2} \\
&= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \overline{Z})}{\sum \hat{\pi}_1^2(Z_i - \overline{Z})^2} \\
&= \frac{1}{\hat{\pi}_1}\frac{\sum(Y_i - \bar{Y})(Z_i - \overline{Z})}{\sum(Z_i - \overline{Z})^2} \\
&= \frac{\sum(Z_i - \overline{Z})^2}{\sum(X_i - \overline{X})(Z_i - \overline{Z})} \times \frac{\sum(Y_i - \bar{Y})(Z_i - \overline{Z})}{\sum(Z_i - \overline{Z})^2}
\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- Which gives the instrumental variable estimator

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})} = \frac{s_{ZY}}{s_{ZX}}$$

- The TSLS estimator of β_1 is the ratio of *the sample covariance between* Z *and* Y to *the sample covariance between* Z *and* X.
- If $Z_i = X_i$, then

$$\hat{\beta}_{2SLS} = \hat{\beta}_{ols}$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$E[\hat{\beta}_{2SLS}] = E\left[\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right]$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0 + \beta_1 X_i + u_i) - (\beta_0 + \beta_1 \bar{X} + \bar{u})](Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0 + \beta_1 X_i + u_i) - (\beta_0 + \beta_1 \bar{X} + \bar{u})](Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right] \\
&= E\left[\frac{\sum \beta_1 (X_i - \bar{X})(Z_i - \bar{Z}) + \sum(u_i - \bar{u})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1+E\left[\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1+E\left[\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1+E\left[\frac{\sum u_i(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Because instrument exogeneity implies $Cov(Z_i, u_i) = 0$,but not $E[u_i|Z_i, X_i] = 0$,then

$$E\left[\frac{\sum u_i(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right] = E\left[\frac{\sum E[u_i|X_i, Z_i](Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\right] \neq 0$$

- Then we have

$$E[\hat{\beta}_{2SLS}] \neq \beta_1$$

- It means that 2SLS estimator is **biased**.

Statistical propertise of 2SLS estimator: Consistent

- We have a simple regression $Y_i = \beta_0 + \beta_1 X_i + u_i$ and take a covariance of Y_i and Z_i

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

- Thus if the instrument is valid,

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

- The population coefficient is the ratio of *the population covariance between Z and Y* to *the popualtion covariance between Z and X*.

Statistical propertise of 2SLS estimator: Consistent

- As discussed in Section 3.7,the sample covariance is a consistent estimator of the population covariance, thus $s_{ZY} \xrightarrow{p} Cov(Z_i, Y_i)$ and $s_{ZX} \xrightarrow{p} Cov(Z_i, X_i)$
- Then the TSLS estimator is **consistent**.

$$\hat{\beta}_{2SLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} = \beta_1$$

Statistical propertise of 2SLS : sampling distribution

- Similar to the expression for the OLS estimator in Equation (4.30,page 183 in S.W)

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}$$

Statistical propertise of 2SLS : sampling distribution

- Similar to the expression for the OLS estimator in Equation (4.30,page 183 in S.W)

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}$$
$$= \frac{\sum[(\beta_0 + \beta_1 X_i + u_i) - (\beta_0 + \beta_1 \bar{X} + \bar{u})](Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}$$

Statistical propertise of 2SLS : sampling distribution

- Similar to the expression for the OLS estimator in Equation (4.30,page 183 in S.W)

$$\begin{aligned}\hat{\beta}_{2SLS} &= \frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\\ &= \frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\\ &= \frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\end{aligned}$$

Statistical propertise of 2SLS : sampling distribution

- Similar to the expression for the OLS estimator in Equation (4.30,page 183 in S.W)

$$\begin{aligned}\hat{\beta}_{2SLS} &= \frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\\ &= \frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\\ &= \frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\\ &= \beta_1+\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\end{aligned}$$

Statistical propertise of 2SLS : sampling distribution

- Similar to the expression for the OLS estimator in Equation (4.30,page 183 in S.W)

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})} \\
&= \frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})} \\
&= \frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})} \\
&= \beta_1+\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})} \\
&= \beta_1+\frac{\frac{1}{n}\sum u_i(Z_i-\bar{Z})}{\frac{1}{n}\sum(X_i-\bar{X})(Z_i-\bar{Z})}
\end{aligned}
$$

Statistical propertise of 2SLS: sampling distribution

- Large sample: $\bar{Z} \cong \mu_z$. Let $q_i = (Z_i - \mu_Z)u_i$,then the numerator

$$\frac{1}{n}\sum u_i(Z_i - \bar{Z}) \cong \frac{1}{n}\sum q_i = \bar{q}$$

- Because $Cov(Z_i, u_i) = 0$ and $E(u_i)$=0,so

$$Cov(Z_i - \mu_Z, u_i) = E[(Z_i - \mu_Z)u_i] = E(q_i) = 0$$

Statistical propertise of 2SLS: sampling distribution

- In addition,the variance of q_i is $\sigma_q^2 = Var[(Z_i - \mu_Z)u_i]$.
- We also have

$$Var(\bar{q}) = \sigma_{\bar{q}}^2 = \frac{\sigma_q^2}{n} = \frac{1}{n}Var[(Z_i - \mu_Z)u_i]$$

- By the C.L.T.(central limit theorem) in large sample,

$$\frac{\bar{q}}{\sigma_{\bar{q}}^2} \xrightarrow{d} N(0,1)$$

Statistical propertise of 2SLS: sampling distribution

- Because the sample covariance is consistent for the population covariance,thus $s_{XY} \xrightarrow{p} Cov(X_i, Y_i)$, then we obtain

$$\hat{\beta}_{2SLS} \cong \beta_1 + \frac{\bar{q}}{Cov(Z_i, Y_i)}$$

- In addition,because $\bar{q} \xrightarrow{d} N(0, \sigma^2_{\bar{q}})$,then we have

$$\frac{\bar{q}}{Cov(Z_i, X_i)} \xrightarrow{d} N(0, \frac{\sigma^2_{\bar{q}}}{[Cov(Z_i, X_i)]^2})$$

Statistical propertise of 2SLS: sampling distribution

- At last, so in large samples $\hat{\beta}_{2SLS}$ is approximately distributed

$$\hat{\beta}_{2SLS} \xrightarrow{d} N(\beta, \sigma^2_{\hat{\beta}_{2SLS}})$$

- Where

$$\sigma^2_{\hat{\beta}_{2SLS}} = \frac{\sigma^2_{\bar{q}}}{[Cov(Z_i, X_i)]^2} = \frac{1}{n}\frac{Var[(Z_i - \mu_Z)u_i]}{Cov[(Z_i, X_i)]^2} \tag{12.8}$$

Statistical propertise of 2SLS: Statistical Inference

- The variance $\hat{\beta}_{2SLS}$ can be estimated by estimating the variance and covariance terms appearing in Equation （12.8),thus

$$SE(\hat{\beta}_{2SLS}) = \sqrt{\frac{\frac{1}{n}\sum(Z_i-\mu_Z)^2\hat{u}_i^2}{n(\frac{1}{n}\sum(Z_i-\mu_Z)X_i)^2}}$$

- Then the square root of the estimate of $\sigma^2_{\hat{\beta}_{2SLS}}$, thus *the standard error of the IV estimator*, which is a little bit complicated. Fortunately,this is done automatically in TSLS regression commands in econometric software packages.
- Because $\hat{\beta}_{2SLS}$ is normally distributed in large samples, hypothesis tests about β can be performed by computing *the t-statistic*,and a 95% large-sample *confidence interval* is given by

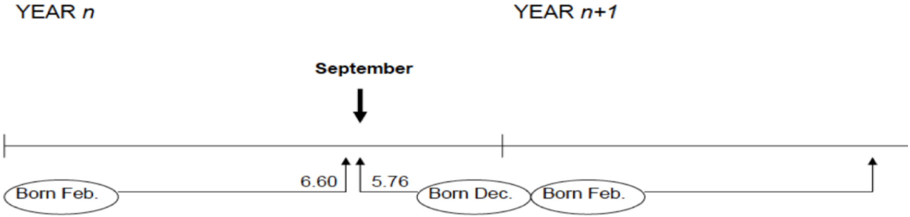
$$\hat{\beta}_{2SLS} \pm 1.96SE(\hat{\beta}_{2SLS})$$

Application: Angrist and Krueger(1991)

- Angrist, Joshua D. and Alan B. Krueger. 1991. “Does Compulsory School Attendance Affect Schooling and Earnings?” The Quarterly Journal of Economics 106 (4):pp979–1014.
- They use **quarter of birth** as an instrument for education to estimate the returns to schooling.

Application: Angrist and Krueger(1991)

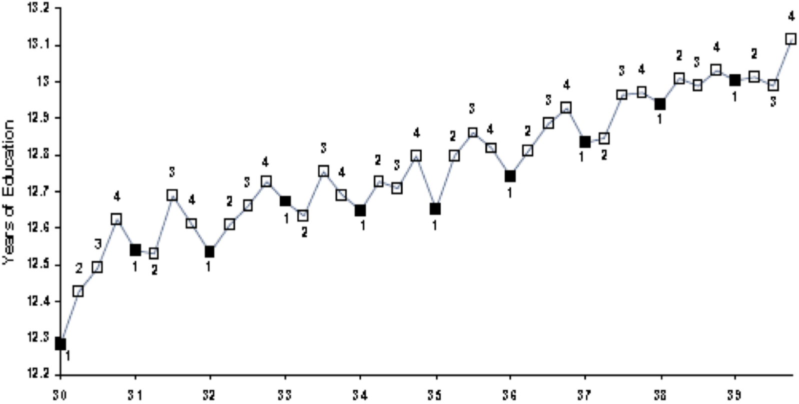
- Why is the Quarter of Birth?
  - In most of the U.S. must attend school *until age 16* (at least during 1938-1967)
  - Age when starting school depends on birthday, so grade when can legally drop out depends on birthday by compulsory schooling laws.



Application: Angrist and Krueger(1991)

- Is Schooling related to Quarter of Birth?(Assumption 1)

A. Average Education by Quarter of Birth (first stage)



Angrist and Krueger(1991): The First Stage

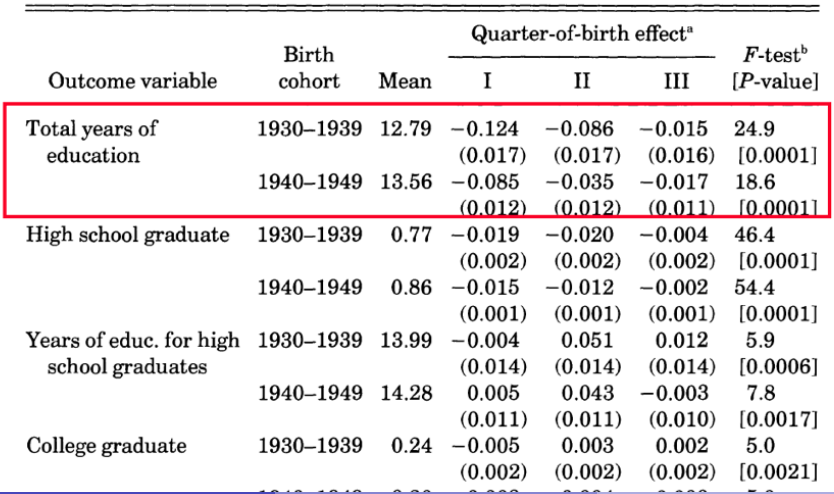
- Does quarter of birth affect education?
- Regress education outcomes on quarter of birth dummy variables:

$$S_{ijc} = \alpha + \beta_1 Q_{1ic} + \beta_2 Q_{2ic} + \beta_3 Q_{3ic} + \epsilon_{ijc}$$

- where individual i, cohort c, education outcome S, birth quarter Q_j
- It is the **first stage** regression

Angrist and Krueger(1991): The First Stage

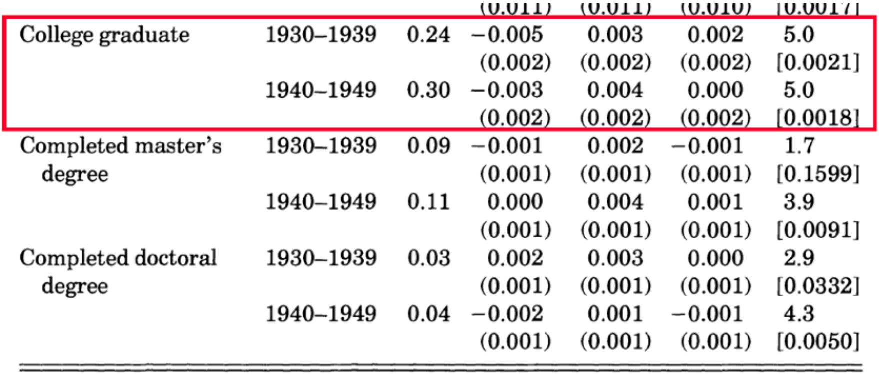
- It shows that Q_j **does** impact education outcomes such as total years of education and high school graduation.



Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124 (0.017)	−0.086 (0.017)	−0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	−0.085 (0.012)	−0.035 (0.012)	−0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	−0.019 (0.002)	−0.020 (0.002)	−0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	−0.015 (0.001)	−0.012 (0.001)	−0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	−0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	−0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]

Angrist and Krueger(1991): exogeneity

- Due to compulsory schooling laws?
- Indirect evidence: on post-secondary outcomes that are not expected to be affected by compulsory schooling laws.

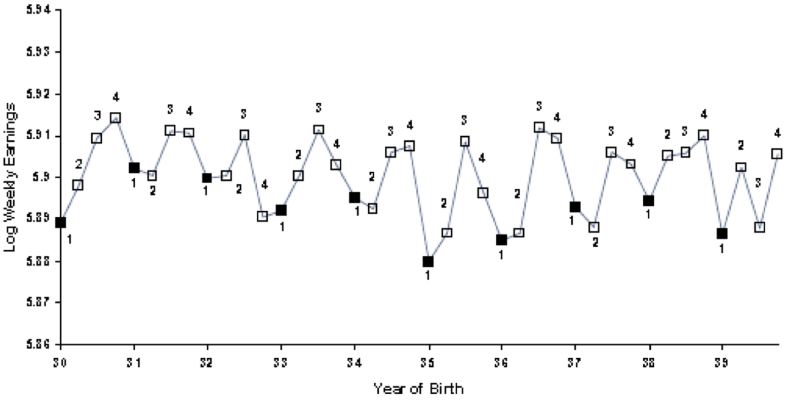


			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]
Completed master's degree	1930–1939	0.09	−0.001	0.002	−0.001	1.7
			(0.001)	(0.001)	(0.001)	[0.1599]
	1940–1949	0.11	0.000	0.004	0.001	3.9
			(0.001)	(0.001)	(0.001)	[0.0091]
Completed doctoral degree	1930–1939	0.03	0.002	0.003	0.000	2.9
			(0.001)	(0.001)	(0.001)	[0.0332]
	1940–1949	0.04	−0.002	0.001	−0.001	4.3
			(0.001)	(0.001)	(0.001)	[0.0050]

Angrist and Krueger(1991): Reduced form

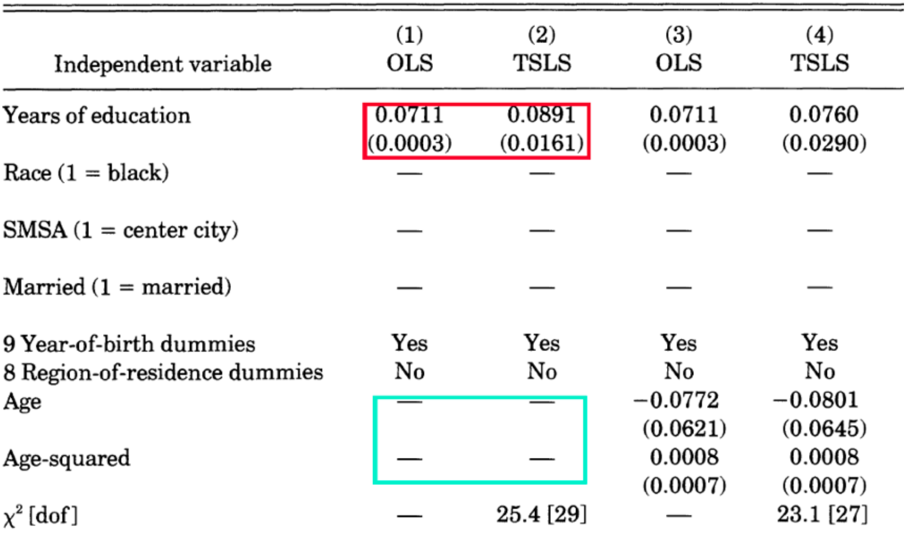
- Is Earnings related to Quarter of Birth?

B. Average Weekly Wage by Quarter of Birth (reduced form)



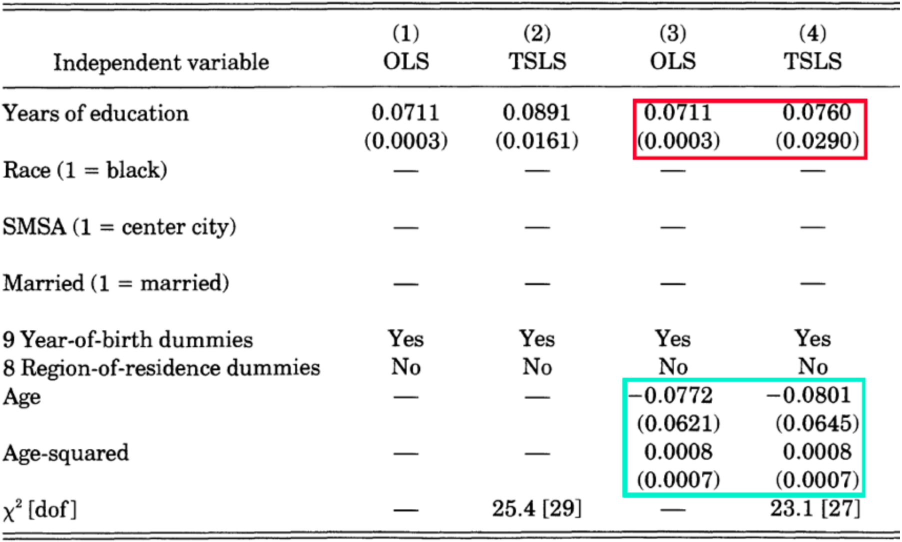
Angrist and Krueger(1991): OLS v.s IV

- IV Estimates



Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760
	(0.0003)	(0.0161)	(0.0003)	(0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	−0.0772	−0.0801
			(0.0621)	(0.0645)
Age-squared	—	—	0.0008	0.0008
			(0.0007)	(0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

Angrist and Krueger(1991): OLS v.s IV with covariates



Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760
	(0.0003)	(0.0161)	(0.0003)	(0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	−0.0772	−0.0801
			(0.0621)	(0.0645)
Age-squared	—	—	0.0008	0.0008
			(0.0007)	(0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

Checking Instrument Validity

Assumption #1 Instrument Relevance

- Instrumental strategy that seems very robust.
- But how to understand that Angrist and Krueger(1991) IV's result larger than that of OLS?
- Bound et al(1995) prove that when instruments have limited explanatory power over endogenous variable,

  1.IV is biased towards OLS in finite samples. 2.May happen even on very large sample

Assumption #1 Instrument Relevance

- Recall 2SLS: a simple OLS regression equation is

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

- Get the predict value from the first stage

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

- Running the second stage regression

$$Y_i = \beta_0 + \beta_1 \widehat{X}_i + u_i$$

- So following the OLS formula in large sample, we can obtain

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{Cov(\widehat{X}, u)}{Var(\widehat{X})}$$

Assumption #1 Instrument Relevance

- An 2SLS version of OVB

$$
\begin{aligned}
\hat{\beta}_{2SLS} &\xrightarrow{p} \beta + \frac{Cov(\hat{X}, u)}{Var(\hat{X})} \\
&= \beta + \frac{Cov(\hat{\pi}_0 + \hat{\pi}_1 Z, u)}{Var(\hat{\pi}_0 + \hat{\pi}_1 Z)} \\
&= \beta + \frac{\hat{\pi}_1 Cov(Z, u)}{\hat{\pi}_1^2 Var(\hat{Z})} \\
&= \beta + \frac{Var(Z)}{Cov(Z, X)} \frac{Cov(Z, u)}{Var(Z)} \\
&= \beta + \frac{Cov(Z, u)}{Cov(Z, X)}
\end{aligned}
$$

Weak Instruments

- Assumption 1: Instrument Relevance

$$Cov(X_i, Z_i) \neq 0$$

.

- Intuition: the more the variation in X is explained by the instruments, thus the more information is available for use in IV regression
- On the contrary, instruments explain little of variation in X are called **Weak Instruments**, thus there is a very weak correlation between X(endogenous variable) and Z(IV).
- Because

$$\hat{\beta}_{2SLS} \xrightarrow{p} \beta + \frac{Cov(Z, u)}{Cov(Z, X)}$$

- So if $Cov(Z, X) = 0$,thus X and Z is *irrelevant*,the bias will approximate to infinity.

Weak Instruments: How to test weak instruments ?

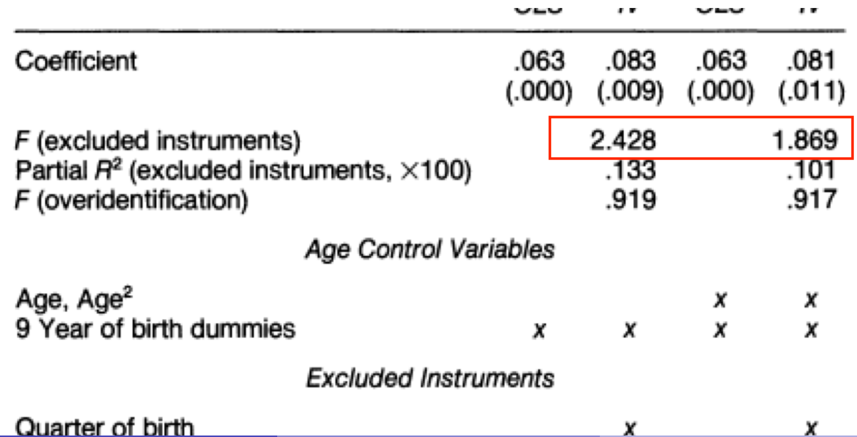
- We should therefore always check whether an instrument is relevant enough.
- Compute the first stage F-statistic provide a measure of the in formation content contained in the instruments.
- Stock and Yogo(2005) showed that

$$E(\beta_{2SLS})-\beta \cong \frac{E(\beta_{ols})-\beta}{E(F)-1}$$

- $E(F)$ is the expectation of the first stage F-statistics.And if $E(F)=10$,the bias of 2SLS, relative to the bias of OLS,is approximately $\frac{1}{9}$, which is small enough to be acceptable.
- *A Rule of Thumb*: **if F-statistic exceeds** 10,then don't need worry about too much.

Angrist and Krueger(1991): Why IV over OLS?

- In Angrist and Krueger(1991),despite large samples sizes, the F-statistics for a test of the joint statistical significance of the excluded exogenous variables in the first-stage regression are not over 2.



	OLS	IV	OLS	IV
Coefficient	.063 (.000)	.083 (.009)	.063 (.000)	.081 (.011)
F (excluded instruments)		2.428		1.869
Partial R^2 (excluded instruments, ×100)		.133		.101
F (overidentification)		.919		.917
Age Control Variables				
Age, Age^2			*x*	*x*
9 Year of birth dummies	*x*	*x*	*x*	*x*
Excluded Instruments				
Quarter of birth		*x*		*x*

Wrap up

- If the correlation between the instruments and the endogenous variable is small, then even the enormous sample sizes do not guarantee that quantitatively important finite sample biases will be eliminated from IV estimates.
- The first assumption of IV method, thus relevance of IV, can be justified by the F-statistic in the first stage.
- Potential Solutions
  - If you have many IVs, some are strong, some are weak. Then discard weak ones.
  - If you only have an weak IV, then find other more stronger IV(easy to say, very hard to do)
  - Employing other estimator(LIML) other than 2SLS methods.

Assumption #2 Instrument Exogeneity

- If the instruments are not exogenous, then TSLS is inconsistent.
- After all, the idea of instrumental variables regression is that the instrument contains information about variation in X_i that is unrelated to the error term u_i.
- *Can we statistically test the assumption that the instruments are exogenous?*
- Answer: In most case,**NO**.
- Assessing whether the instruments are exogenous necessarily requires making an expert judgment based on personal knowledge and expert opinion of the application.(“讲好故事”)
- In some case,you can test partially,thus **overidentification test**.

Assumption #2 Instrument Exogeneity

- Terminology: The relationship between the number of instruments(m) and the number of endogenous regressors(k)
    - **exactly(just) identified**:$m = k$
    - **overidentified** $m > k$
    - **underidentified** $m < k$
- when the coefficients are just identified, you can't do a formal statistical test of the hypothesis that the instruments are in fact exogenous.
- If, however, there are more instruments than endogenous regressors, then there is a statistical tool that can be helpful in this process: the so-called test of *overidentifying restrictions*.

Overidentification-test:Intuition

- Suppose there are two valid instruments: Z_1 Z_2(you are very lucky.)
- Then you could compute two separate TSLS estimates.
- Intuitively,if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *overidentifying restrictions test* makes this comparison in a statistically precise way.

Overidentification test:

- Our model is a multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + \beta_{k+1} W_{1,i} + ... + \beta_{k+r} W_{r,i} + u_i \tag{12.13}$$

- Where
  - Y_i is the *dependent variable*
  - $X_1, X_2, ...X_k$ are K *endogenous regressors*
  - $W_1, X_2, ...W_r$ are the *additional exogenous variables*
  - we have m instruments,$Z_1, Z_2, ...Z_m$,*instrumental variables*
  - u_i is the regression error term.

Overidentification test:

- A set of m instruments, $Z_1, Z_2, ...Z_m$
- then 2sls regression

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1,i} + \beta_2 \hat{X}_{2,i} + ... + \beta_k \hat{X}_{k,i} + \beta_{k+1} W_{1,i} + ... + \beta_{k+r} W_{r,i} + u_i \quad (12.13)$$

- then we can get the predict value of $\hat{u_i}^{TSLS}$

Overidentification test:

- Let

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

- Let F denote the homoskedasticity-only F-statistic testing the hypothesis that $\delta_0 = ... = \delta_m = 0$
- Then the overidentifying restrictions test statistic is $J = mF$
- Under the null hypothesis that all the instruments are exogenous,

$$J \xrightarrow{d} \chi^2_{m-k}$$

- Where $m - k$ is the "degree of over-identification," that is, the number of instruments minus the number of endogenous regressors.

Application: Demand for Cigarettes

- A serious public health issue: huge externalities
- One policy tool is to tax cigarettes so heavily that current smokers cut back and potential new smokers are discouraged from taking up the habit.
- Precisely how big a tax hike is needed to make a dent in cigarette consumption?
- For example, what would the after-tax sales price of cigarettes need to be to achieve a 20% reduction in cigarette consumption?
- The answer to this question depends on the **elasticity of demand** for cigarettes.

Application: Demand for Cigarettes

- Because of the interactions between supply and demand, the elasticity of demand for cigarettes cannot be estimated consistently by an OLS regression of log quantity on log price.
- Using annual data for the 48 contiguous U.S. states for in 1995,we therefore use TSLS to estimate the elasticity of demand for cigarettes.
- The instrumental variable, $SalesTax_i$, is the portion of the tax on cigarettes arising from the general sales tax,measured in dollars per pack.
- Cigarette consumption, $Q_i^{cigarettes}$, is the number of packs of cigarettes sold per capita in the state,
- and the price, $P_i^{cigarettes}$,is the average real price per pack of cigarettes including all taxes.

Application: Demand for Cigarettes

- We consider quantity and price changes that occur over 10-year periods.
- Dependent variable:

$$\Delta ln(Q_i^{cigarettes}) = ln(Q_{i,1995}^{cigarettes}) - ln(Q_{i,1985}^{cigarettes})$$

- Independent variable:

$$\Delta ln(P_i^{cigarettes}) = ln(P_{i,1995}^{cigarettes}) - ln(P_{i,1985}^{cigarettes})$$

- Control variable:

$$\Delta ln(Inc_i) = ln(Inc_{i,1995}) - ln(Inc_{i,1985})$$

Application: Demand for Cigarettes

- Two instruments

1. the change in the sales tax over 10 years,

$$\Delta SalesTax_i = SalesTax_{i,1995} - SalesTax_{i,1985}$$

2. the change in the cigarette-specific tax over 10 years

$$\Delta CigTax_i = CigTax_{i,1995} - CigTax_{i,1985}$$

Application: Demand for Cigarettes

- The first stage

$$\widehat{\ln(P_{i,1995}^{cigarettes}) - \ln(P_{i,1985}^{cigarettes})} = \underset{(0.03)}{0.53} - \underset{(0.22)}{0.22}[\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})]$$
$$+ \underset{(0.0044)}{0.0255}(SalesTax_{i,1995} - SalesTax_{i,1985}). \qquad (12.18)$$

Application: Demand for Cigarettes

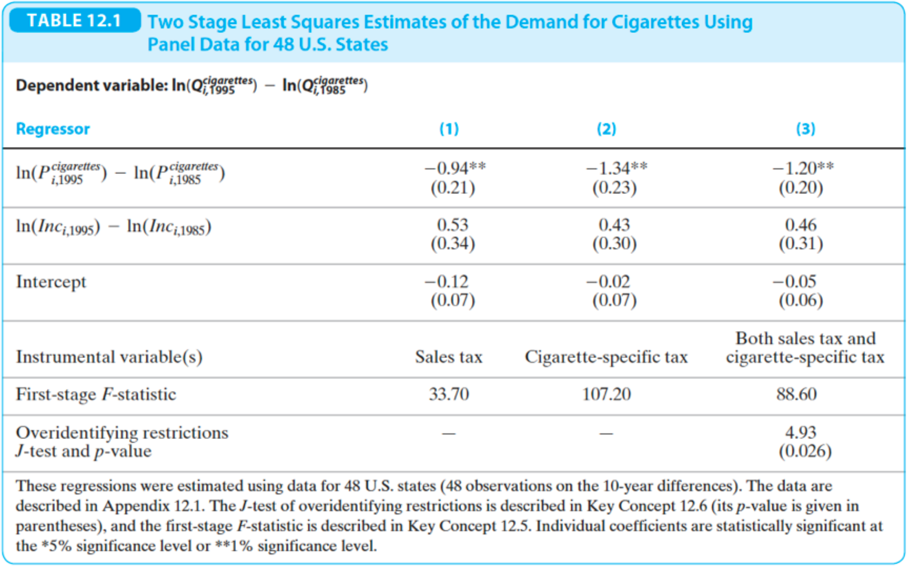


TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

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)	−1.34** (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% significance level or **1% significance level.

Application: Demand for Cigarettes

- Over-identifying J-test **reject** the null hypothesis that both the instruments are exogenous at the 5% significant level$(p - value = 0.026)$

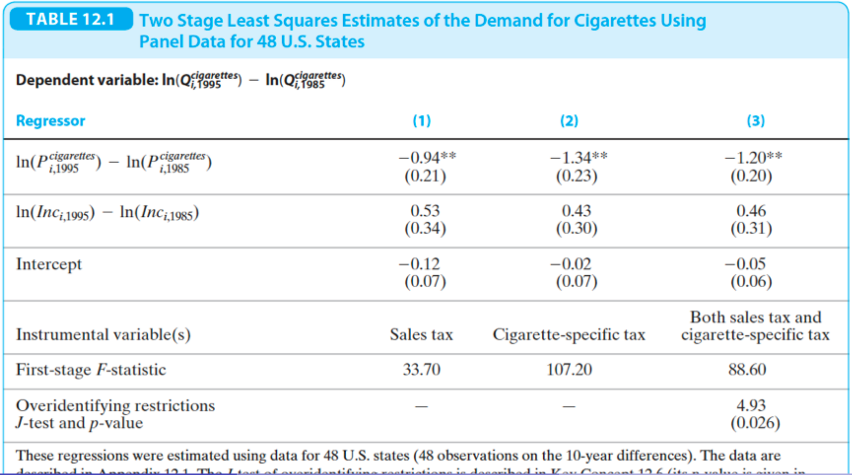


TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

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

Application: Demand for Cigarettes

- The reason the J-statistic rejects the null hypothesis that both instruments are exogenous is that the two instruments produce rather different estimated coefficients.
- The J-statistic rejection means that the regression in column (3) is based on invalid instruments (the instrument exogeneity condition fails).

Application: Demand for Cigarettes

- The J-statistic rejection says that at least one of the instruments is endogenous, so there are three logical possibilities
  - The sales tax is exogenous but the cigarette-specific tax is not, in which case the column (1) regression is reliable;
  - the cigarette-specific tax is exogenous but the sales tax is not, so the column (2) regression is reliable;
  - or neither tax is exogenous, so neither regression is reliable. The statistical evidence cannot tell us which possibility is correct, so we must use our judgement.

Application: Demand for Cigarettes

- **We think** that the case for the exogeneity of the general sales tax is stronger than that for the cigarette-specific tax.
- because the political process can link changes in the cigarette-specific tax to changes in the cigarette market and smoking policy.
- if smoking decreases in a state because it falls out of fashion, there will be fewer smokers and a weakened lobby against cigarettespecific tax increases, which in turn could lead to higher cigarette-specific taxes.

Application: Demand for Cigarettes

- So the result that use the cigarette-only tax as an instrument and adopting the price elasticity estimated using the general sales tax as an instrument is more reliable.
- The estimate of -0.94 indicates that cigarette consumption is somewhat elastic:An increase in price of 1% leads to a decrease in consumption of 0.94%.

Instrumental Variable for multiple regression

IV for multiple regression(Key Concept 12.1)

- Our model is a multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + \beta_{k+1} W_{1,i} + ... + \beta_{k+r} W_{r,i} + u_i \tag{12.13}$$

- Where
  - Y_i is the *dependent variable*
  - $X_1, X_2, ...X_k$ are K *endogenous regressors*
  - $W_1, X_2, ...W_r$ are the *additional exogenous variables*
  - we have m instruments,$Z_1, Z_2, ...Z_m$,*instrumental variables*
  - u_i is the regression error term.

Two Conditions for Valid Instruments

- A set of m instruments ,$Z_1, Z_2, ...Z_m$ must satisfy the following two conditions to be valid:

1. Instrument Relevance:
   - In general,let $\widehat{X}^*_{1i}$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z) and the included exogenous regressors (W), and let "1" denote the constant regressor that takes on the value 1 for all observations. Then $(\widehat{X}^*_{1i}, ..., \widehat{X}^*_{ki}, W_1, X_2, ...W_r, 1)$ are *not perfectly multicollinear*.
   - If there is only one X, then for the previous condition to hold, at least one Z must have a non-zero coefficient in the population regression of X on the Z and the W.
2. Instrument Exogeneity
   - The instruments are uncorrelated with the error term,

$$Cov(Z_{1i}, u_i) = 0, ..., Cov(Z_{mi}, u_i) = 0$$

The IV Regression Assumptions(Key Concept 12.4)

- The variables and errors in the IV regression model in Key Concept 12.1 satisfy the following:
  - 1. $E(ui|W_{1i},...,W_{ri}) = 0$
  - 2. $(X_{1i},...,X_{ki},W_{1i},...,W_{ri},Z_{1i},...,Z_{mi},Yi)$ are i.i.d. draws from their joint distribution;
  - 3. Large outliers are unlikely: The X,W,Z, and Y have nonzero finite fourth moments;
  - 4. The two conditions for a valid instrument hold.
- Under the IV regression assumptions,the TSLS estimator is consistent and normally distributed in large samples.
- Because the sampling distribution of the TSLS estimator is normal in large samples,the general procedures for statistical inference (hypothesis tests and confidence intervals) in regression models extend to TSLS regression.

Review the last lecture

Instrument Variables:Constant-effect

- Instrumental Variable is a useful method to make causal inference. It can eliminate
  - Omitted Variable Bias
  - Measurement Error
  - Reverse Causality
- Two Assumptions
  - Relevance(Weak Instrument): It can be test by the first stage regression and F-statistic.
  - Exogeneity: Can't be test formally but argue it using professional knowledges.
- Estimation and Inference
  - When IV satisify these two assumptions,the causal effect of coefficients of interest,TSLS estimator,β_{TSLS} can be NOT unbiased but **consistent**.
  - The sampling distribution of the TSLS estimator is also normal in large

IV with Heterogeneous Causal Effects: Simple Case

Example: Angrist(1990)

- Topic: How does **veteran** status effect on earnings
- Methods: Instrumental Variable
- Use the lottery outcome as an instrument for veteran status

Example: Angrist(1990) Background

- In the 1960s and 70s young men in the US were at risk of being drafted for military service in Vietnam.
- Fairness concerns led to the institution of a draft lottery in 1970 that was used to determine priority for conscription.
- In each year from 1970 to 1972, random sequence numbers were randomly assigned to each birth date in cohorts of 19-year-olds.
    - Men with lottery numbers below a cutoff were eligible for the draft
    - Men with lottery numbers above the cutoff were not.

Example: Angrist(1990) Instrumental Variables

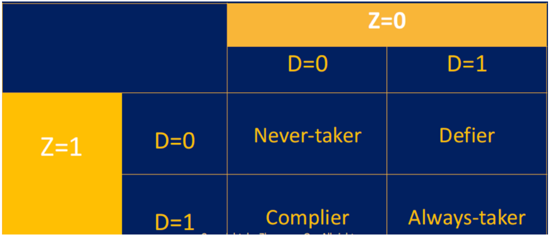
- The instrument(Z_i) is thus defined as follows:
  - $Zi = 1$ if lottery implied individual i would be draft eligible,
  - $Zi = 0$ if lottery implied individual i would NOT be draft eligible.
- The econometrician observes treatment status(D_i) as follows:
  - $Di = 1$ if individual i served in the Vietnam war (veteran)
  - $Di = 0$ if individual i did not serve in the Vietnam war (not veteran)

Example: Angrist(1990): IV's Relevance and Exogenous

- While the lottery didn't completely determine veteran status, it certainly mattered: relevance.
- The lottery outcome was random and seems reasonable to suppose that its only effect was on veteran status: exogenous.

Example: Angrist(1990): heterogeneous effects

- We can classify individuals according to assignment(Z) an treatment(X) into four parts



		Z=0	
		D=0	D=1
Z=1	D=0	Never-taker	Defier
	D=1	Complier	Always-taker

Local Average Treatment Effect(LATE)

- So IV estimate only get the X effect on Y on the subpopulation-compilers.
- Angrist and Imbens(1994) called it as **Local Average Treatment Effect(LATE)**, thus the treatment effect on those that change their behavior under the instrument.

IV with Heterogeneous Causal Effects: Generalization

Introduction

- If the population is *heterogeneous*, then the i^{th} individual now has his or her own causal effect, β_{1i},then the population regression equation can be written
$$Y_i = \beta_{0i} + \beta_{1i}X_i + u_i \tag{13.9}$$
- β_{1i} is a random variable that,just like u_i, reflects unobserved variation across individuals.
- The average causal effect is the population mean value of the causal effect,$E(\beta_{1i})$ which is the expected causal effect of a randomly selected member of the population.

OLS with Heterogeneous Causal Effects

- If there is heterogeneity in the causal effect and if X_i is randomly assigned, then the differences estimator is a consistent estimator of the average causal effect.

$$\begin{aligned}\hat{\beta}_{ols} = \frac{s_{XY}}{s_X^2} \xrightarrow{p} \frac{Cov(Y_i, X_i)}{Var(X_i)} &= \frac{Cov(\beta_{0i} + \beta_{1i}X_i + u_i, X_i)}{Var(X_i)} \\ &= \frac{Cov(\beta_{1i}X_i, X_i)}{Var(X_i)} \\ &= E(\beta_{1i})\end{aligned}$$

- Thus, if X_i is randomly assigned, $\hat{\beta}_1$ is a *consistent* estimator of the average causal effect $E(\beta_{1i})$.

IV Regression with Heterogeneous Causal Effects

- Specifically, suppose that X_i is related to Z_i by the linear model

$$X_i = \pi_{0i} + \pi_{1i} Z_i + v_i$$

- where the coefficients π_{0i} and π_{1i} vary from one individual to the next. And it is the first-stage equation of TSLS with the modification of heterogeneous effect of Z on X.

IV Regression with Heterogeneous Causal Effects

- Then TSLS estimator becomes

$$\hat{\beta}_{2SLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}$$

- **Excercise**: *prove it by yourself (refers to Appendix 13.2)*
- The TSLS estimator converges in probability to the ratio of the expected value of the product of β_{1i} and π_{1i} to the expected value of π_{1i}.

IV Regression with Heterogeneous Causal Effects

- It is a **weighted** average of the individual causal effects β_{1i}, The weights are $\frac{\pi_{1i}}{E(\pi_{1i})}$, which measure the relative degree to which the instrument influences whether the i_{th} individual receives treatment,
- In other words,TSLS estimator is a consistent estimator of a *weighted average of the individual causal effects*, where the individuals who receive the *most weight* are those for *whom the instrument is most influential*.

IV Regression with Heterogeneous Causal Effects

- Three special cases:
  - The treatment effect is the same for all individuals.

    $$\beta_{1i} = \beta_1$$

  - The instrument affects each individual equally.

    $$\pi_{1i} = \pi_1$$

  - The heterogeneity in the treatment effect and heterogeneity in the effect of the instrument are uncorrelated.

    $$Cov(\beta_{1i}\pi_{1i}) = 0$$

IV Regression with Heterogeneous Causal Effects

- LATE equals to the ATE: all three cases we have

$$\frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})} = E(\beta_{1i}) = \beta_1$$

- Aside from these three special cases, in general the local average treatment effect differs from the average treatment effect.

IV Regression with Heterogeneous Causal Effects: Implications

- Different instruments can identify different parameters because they estimate the impact on different populations.
- The difference arises because each researcher is implicitly estimating a different weighted average of the individual causal effects in the population.
- Recall: **J-test of overidentifying restrictions** can reject if the two instruments estimate different local average treatment effects,even if both instruments are valid. In general neither estimator is a consistent estimator of the average causal effect.

In Summary

- The IV paradigm provides a powerful and flexible framework for causal inference.
- An alternative to random assignment with a strong claim on internal validity.
- The LATE framework highlights questions of external validity
    - Can one instrument identify the average effect induced by another source of variation?
    - Can we go from average effects on compliers to average effects on the entire treated population or an unconditional effect?
- The answer to these questions is usually: **NO**, at least without additional assumptions.

Some Practical Guides by Angrist and Pischke(2012)

Practical Guides

1. Check IV relevance
   - Always report the first stage and think about whether it makes sense(Signs and magnitudes)
   - Always report the F-statistic on the excluded instruments. The bigger,the better. Don't forget the rule of thumb.($F > 10$)
2. Check exclusion restriction
   - The exclusion restriction cannot be tested directly, but it can be falsified
   - Run and examine the reduced form(regression of dependent variable on instruments) and look at the coefficients, t-statistics and F-statistics for excluded instruments.
   - Because the reduced form is proportional to the casual effect of interest and is unbiased(OLS), so we should see the causal relation in the reduced form.If you can't see the causal relation in the reduced form,it's probably not there

Practical Guides

3. Provide a substantive explanation for observed difference between 2SLS and OLS
   - How bid is the difference? What does this tell you?
   - Is the coefficient bigger when theory of endogeneity suggests it should be smaller? If so, why?
   - Measurement Error or heterogeneous effects?
4. If you have multiple instruments, report over-identification tests.
   - Pick your best single instrument and report just-identified estimates using this one only because just-identified IV is relatively unlikely to be subject to a weak bias.
   - Worry if it is substantially different from what you get using multiple instruments.
   - Check over-identified 2SLS estimates with LIML. LIML is less than precise than 2SIS but also less biased. If the results come out similar, be happy. If not, worry, and try to find stronger instruments.

How to Evaluate IV paper in a simple way?

1. Relevant: The first stage regression
   - Does the author report the first stage regression?
   - Does the instrument perform well in the first stage?
   - Testable: rule of thumb: first stage $F > 10$
2. Exclusion restriction:
   - Is the instrument exogenous enough?(the random assignment is the best)
   - Would you expect a direct effect of Z on Y
   - Not directly testable: Except when equation is overidentified.
3. What LATE is being estimated?
   - Whose behavior is affected by the instrument?
   - Is this the LATE you would want? Is it a quantify of theoretical interest?
   - Would other LATEs possible yield different estimates?

An good example: Long live Keju

Chen, Kung and MA(2017)

- Title: Long Live Keju! The Persistent Effects of China's Imperial Examination System.
- Topic: Long term persistence of human capital:the effect of **Keju**
- Dependent Variable: education level in 2010
- Indepenet Variable: the density of **jinshi** in the Ming-Qing dynasties
- Data: 272 perfectures in *jinshi*.

Chen, Kung and MA(2018)

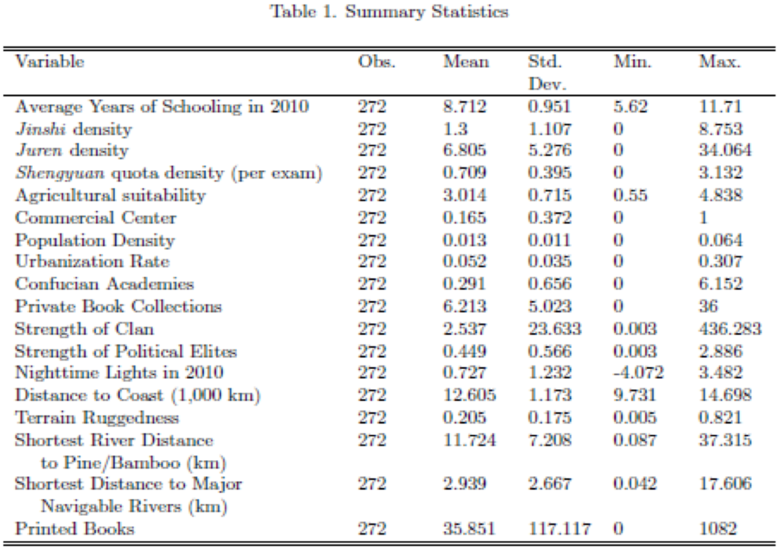


Table 1. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Average Years of Schooling in 2010	272	8.712	0.951	5.62	11.71
Jinshi density	272	1.3	1.107	0	8.753
Juren density	272	6.805	5.276	0	34.064
Shengyuan quota density (per exam)	272	0.709	0.395	0	3.132
Agricultural suitability	272	3.014	0.715	0.55	4.838
Commercial Center	272	0.165	0.372	0	1
Population Density	272	0.013	0.011	0	0.064
Urbanization Rate	272	0.052	0.035	0	0.307
Confucian Academies	272	0.291	0.656	0	6.152
Private Book Collections	272	6.213	5.023	0	36
Strength of Clan	272	2.537	23.633	0.003	436.283
Strength of Political Elites	272	0.449	0.566	0.003	2.886
Nighttime Lights in 2010	272	0.727	1.232	-4.072	3.482
Distance to Coast (1,000 km)	272	12.605	1.173	9.731	14.698
Terrain Ruggedness	272	0.205	0.175	0.005	0.821
Shortest River Distance to Pine/Bamboo (km)	272	11.724	7.208	0.087	37.315
Shortest Distance to Major Navigable Rivers (km)	272	2.939	2.667	0.042	17.606
Printed Books	272	35.851	117.117	0	1082

Chen, Kung and MA(2018)

- The effect of Keju on human capital at present
- Run regression

$$lnY_i = \beta ln(Keju_i) + \gamma_1 X_i^c + \gamma_2 X_i^h + \alpha_p + u_i$$

- Y_i: 2010 年 i 地区的平均受教育年限。
- $Keju_i$: 明清时期 i 地区获得进士的人数。
- X_i^c: 控制变量（当代），包括经济繁荣程度 (夜间灯光)；地理因素：该地区到海选距离、地形（免于遭受自然灾害）。
- X_i^c: 控制变量（历史）：
  - 历史经济繁荣程度
  - 基础教育设施
  - 社会和政治影响力

Chen, Kung and MA(2018): OLS

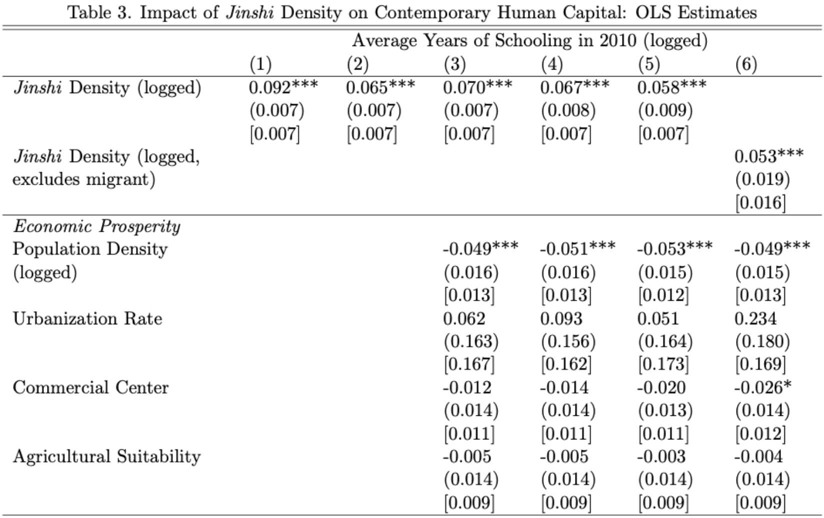


Table 3. Impact of *Jinshi* Density on Contemporary Human Capital: OLS Estimates

	Average Years of Schooling in 2010 (logged)					
	(1)	(2)	(3)	(4)	(5)	(6)
Jinshi Density (logged)	0.092***	0.065***	0.070***	0.067***	0.058***	
	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)	
	[0.007]	[0.007]	[0.007]	[0.007]	[0.007]	
Jinshi Density (logged, excludes migrant)						0.053***
						(0.019)
						[0.016]
Economic Prosperity						
Population Density (logged)			-0.049***	-0.051***	-0.053***	-0.049***
			(0.016)	(0.016)	(0.015)	(0.015)
			[0.013]	[0.013]	[0.012]	[0.013]
Urbanization Rate			0.062	0.093	0.051	0.234
			(0.163)	(0.156)	(0.164)	(0.180)
			[0.167]	[0.162]	[0.173]	[0.169]
Commercial Center			-0.012	-0.014	-0.020	-0.026*
			(0.014)	(0.014)	(0.013)	(0.014)
			[0.011]	[0.011]	[0.011]	[0.012]
Agricultural Suitability			-0.005	-0.005	-0.003	-0.004
			(0.014)	(0.014)	(0.014)	(0.014)
			[0.009]	[0.009]	[0.009]	[0.009]

Chen, Kung and MA(2018): Potential Bias

- OVB: that are simultaneously associated with both historical jinshi density and years of schooling today.
- For instance, prefectures that had produced more jinshi may be associated with unobserved (natural or genetic) endowments.

Chen, Kung and MA(2018): Instrumental Variable

- IV: Distance to the Printing Ingredients (Pine and Bamboo) as the Instrumental Variable of Keju
- A logic chain:

$$\begin{aligned} More;\ suceeded\ in\ Keju &\Longleftrightarrow more\ references\ books \\ &\Longleftrightarrow print\ references\ books\ in\ centers \\ &\Longleftrightarrow print\ centers\ locates\ nearby\ some\ ingr \end{aligned}$$

Chen, Kung and MA(2018): First Stage

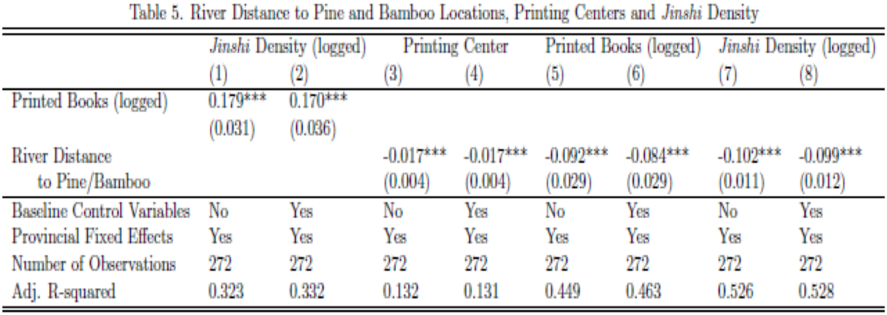


Table 5. River Distance to Pine and Bamboo Locations, Printing Centers and *Jinshi* Density

	Jinshi Density (logged)		Printing Center		Printed Books (logged)		*Jinshi* Density (logged)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Printed Books (logged)	0.179***	0.170***						
	(0.031)	(0.036)						
River Distance to Pine/Bamboo			-0.017***	-0.017***	-0.092***	-0.084***	-0.102***	-0.099***
			(0.004)	(0.004)	(0.029)	(0.029)	(0.011)	(0.012)
Baseline Control Variables	No	Yes	No	Yes	No	Yes	No	Yes
Provincial Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	272	272	272	272	272	272	272	272
Adj. R-squared	0.323	0.332	0.132	0.131	0.449	0.463	0.526	0.528

Chen, Kung and MA(2018): Reduced-form and 2SLS

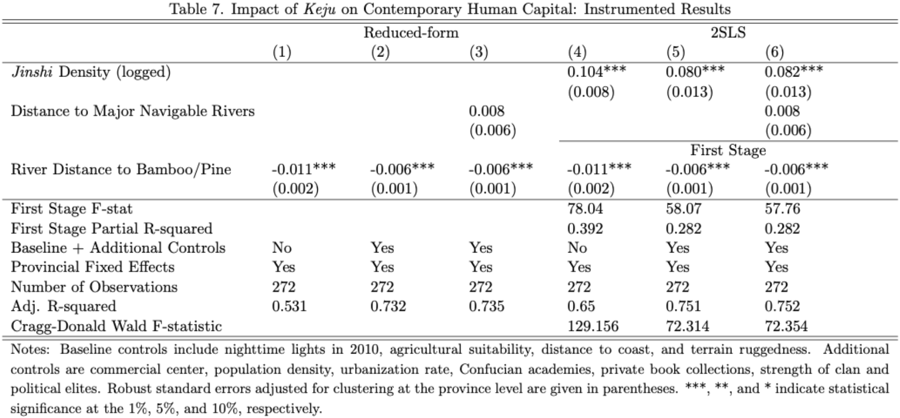


Table 7. Impact of *Keju* on Contemporary Human Capital: Instrumented Results

	Reduced-form			2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
Jinshi Density (logged)				0.104***	0.080***	0.082***
				(0.008)	(0.013)	(0.013)
Distance to Major Navigable Rivers			0.008			0.008
			(0.006)			(0.006)
				First Stage		
River Distance to Bamboo/Pine	-0.011***	-0.006***	-0.006***	-0.011***	-0.006***	-0.006***
	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
First Stage F-stat				78.04	58.07	57.76
First Stage Partial R-squared				0.392	0.282	0.282
Baseline + Additional Controls	No	Yes	Yes	No	Yes	Yes
Provincial Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	272	272	272	272	272	272
Adj. R-squared	0.531	0.732	0.735	0.65	0.751	0.752
Cragg-Donald Wald F-statistic				129.156	72.314	72.354

Notes: Baseline controls include nighttime lights in 2010, agricultural suitability, distance to coast, and terrain ruggedness. Additional controls are commercial center, population density, urbanization rate, Confucian academies, private book collections, strength of clan and political elites. Robust standard errors adjusted for clustering at the province level are given in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10%, respectively.

Where Do Valid Instruments Come From?

Where do we find an IV?

- Generally Speaking
  - “可遇不可求”
- Two main approaches
  1. Economic Theory/Logics
  2. Exogenous Source of Variation in X(natural experiments)

Where do we find an IV?

- Example 1: Does putting criminals in jail reduce crime?
- Run a regression of crime rates(d.v.) on incarceration rates(id.v) by using annual data at a suitable level of jurisdiction(states) and covariates (economic conditions)
- *Simultaneous causality bias*: crime rates goes up, more prisoners and more prisoners,reduced crime.
- IV: it must affect the incarceration rate but be unrelated to any of the unobserved factors that determine the crime rate.
- Levitt (1996) suggested that *lawsuits aimed at reducing prison overcrowding* could serve as an instrumental variable.
- Result: The estimated effect was three times larger than the effect estimated using OLS.

Where do we find an IV?: Class Size and Test Score

- Example 2: Does cutting class sizes increase test scores?
- *Omited Variable bias*: such as parental interest in learning, learning opportunities outside the classroom, quality of the teachers and school facilities.
- IV: correlated with class size (relevance) but uncorrelated with the omitted determinants of test performance.
- Hoxby (2000) suggested biology. Because of random fluctuations in timings of births, the size of the incoming kindergarten class varies from one year to the next.
- But potential enrollment also fluctuates because parents with young children choose to move into an improving school district and out of one in trouble. She used the deviation of potential enrollment from its long-term trend as her instrument.
- Result: the effect on test scores of class size is small

Where do we find an IV?

1. Institutional Background

- Angrist(1990)-draft lottery: Vietnam veterans were randomly designated based on birth day used to estimate the wage impact of a shorter work experience.
- Acemoglu, Johnson, and Robinson(2001): the dead rate of some diseases in some areas to estimate the impact of institutions to economic growth.
- Feng et al.(2012) “The Returns to Education in China: Evidence from the 1986 Compulsory Education Law”.
- Li and Zhang(2007),Liu(2012)- “One Child policy”

Where do we find an IV?

2. Natural conditions(geography,weather,disaster)
   - the Rainfall,Hurricane,Earthquake,Tsunami...
   - the number of Rivers: Hoxby(2000)
   - Ying Bai and Ruixue Jia(2014)-“keju” and “the number of small rivers”

Where do we find an IV?

3. Economic theory and Economic logic
   - study the alcohol consumption and income relationship. alcohol price in a local market may be as a instrument of alcohol consumption.
   - Angrist & Evans(1998): have same sex or different sex children used to estimate the impact of an additional birth on women labor supply.