Lecture 10: Instrumental Variables

Introduction ot Econometrics,Fall 2018

Zhaopeng Qu

Nanjing University

11/22/2018

1. Review Previous Lectures
2. Instrumental Variable Method
3. Checking Instrument Validity
4. Instrumental Variable for multiple regression

Review Previous Lectures

Threatens to Internal Validity

- Three important threats to internal validity 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.
- intuitionally, 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

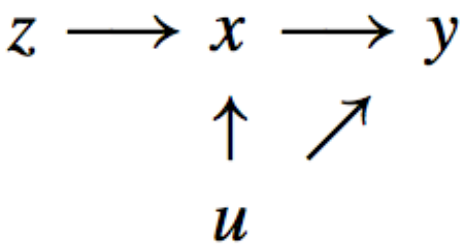
- An instrumental variable Z_i must satisfy the following 2 properties:
  1. **Instrumental relevance**: Z_i should be **correlated** with the casual variable of interest, X_i (endogenous variable),thus

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

.

  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 $hatX_i$ to obtain the Two Stage Least Squares estimator $\hat{\beta}_{2SLS}$

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

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

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

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.

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 popualtion 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 expession 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 numberator

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

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

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

- 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

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

- 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

$$\sigma^2_{\bar{q}} \qquad 1\, Var[(Z_i - \mu_Z)u_i] \qquad (12.8)$$

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

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

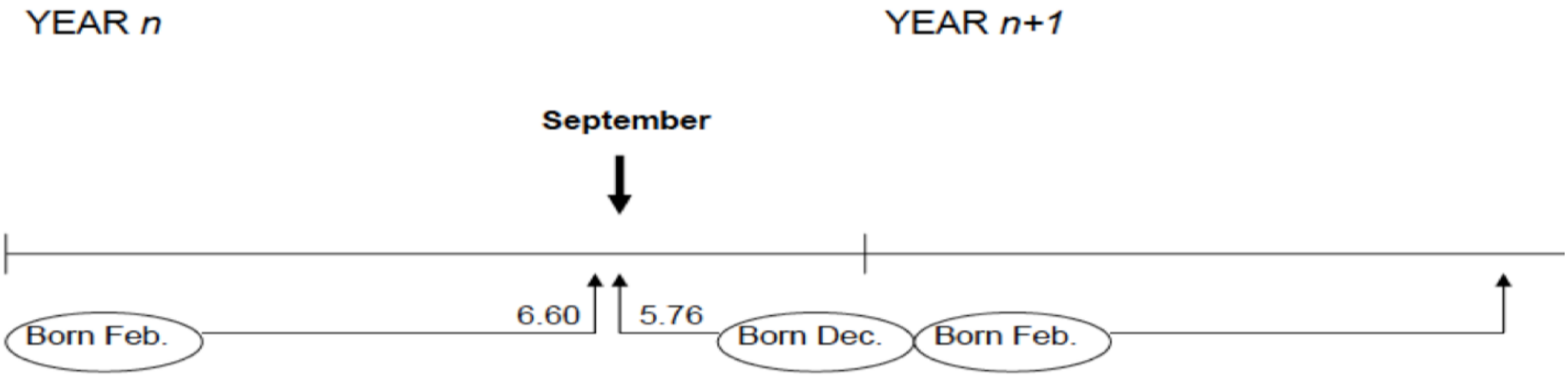
Application to the demand for cigarettes

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): pp. 979–1014.
- They use **quarter of birth** as an instrument for education to estimate the returns to schooling.

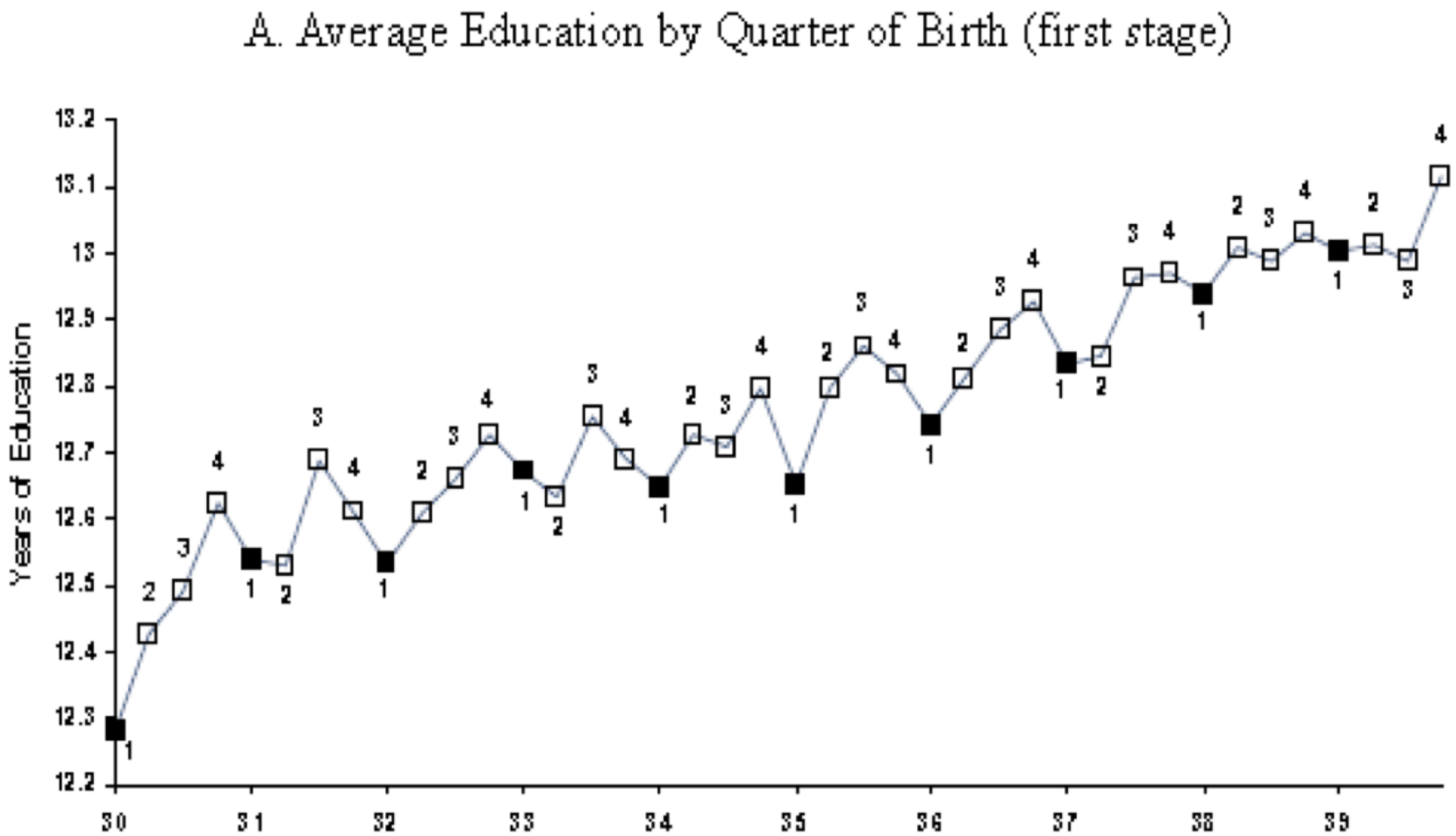
Application: Angrist and Krueger(1991)

- 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)



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	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(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]

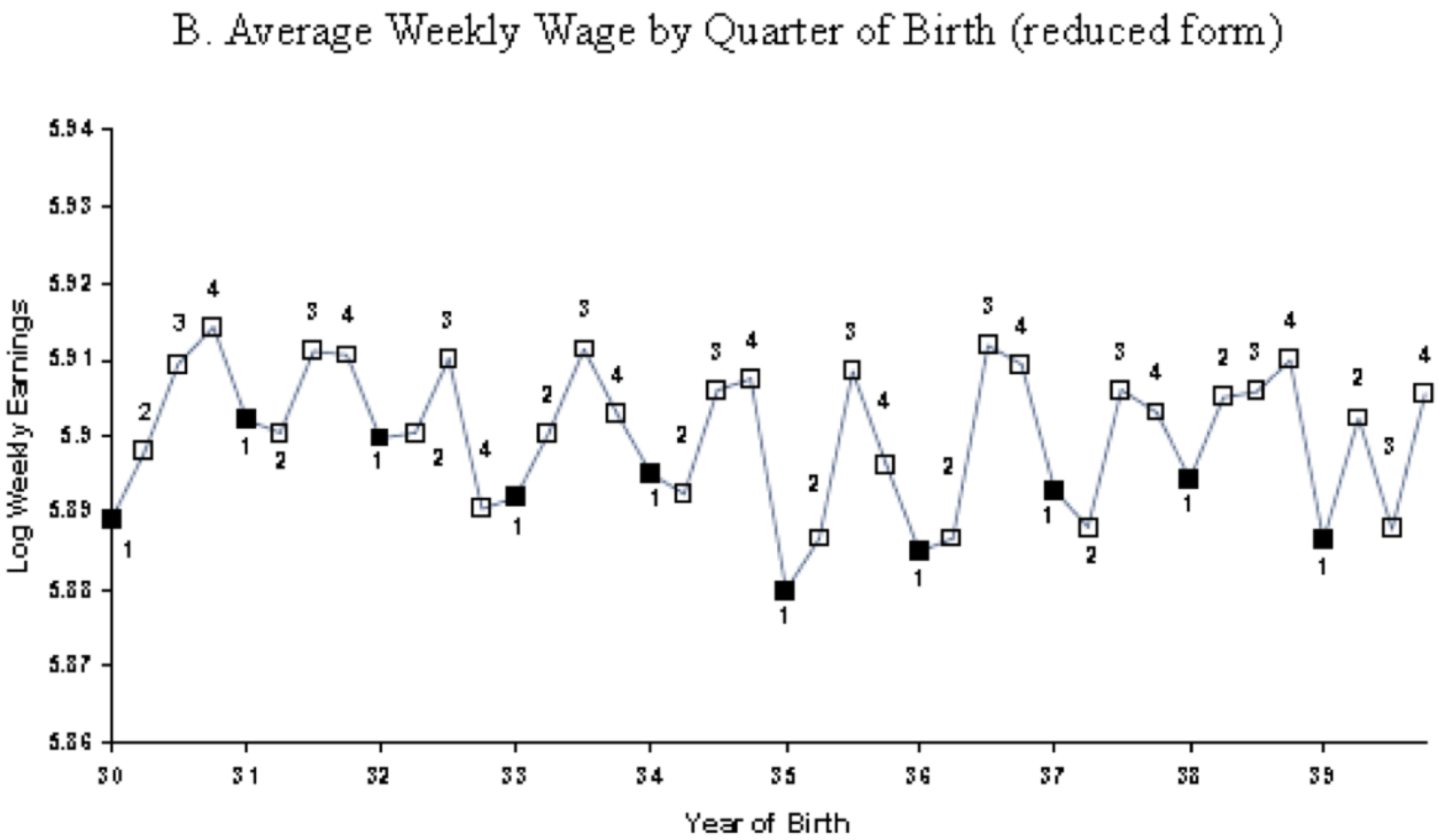
Angrist and Krueger(1991): exogeneity

- 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?



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]

Extension: with control variables

- We can weaken the instrument exogeneity assumption by including area characteristics as control variables
- The Instrumental variables model is extended by including the control variables $W_{1i}, ..., W_{ri}$
- Then the *stuctural equation* is

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \delta_1 W_{1,i} + ... + \delta_r W_{r,i} + u_i$$

- Then the *first stage* is

$$X_i = \pi_0 + \pi_1 Z_{1,i} + \gamma_1 W_{1,i} + ... + \gamma_r W_{r,i} + v_i$$

- The Instrument *exogeneity condition* is now conditional on the included these exogenous controlled variables,$W_{1i}, ..., W_{ri}$

$$Cov(Z_i, u_i | W_{1i}, ..., W_{ri})$$

- Note: control variables $W_{1i}, ..., W_{ri}$ in both regressions are totally

Extension: with control variables

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

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

- Running the second stage regression

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

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

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

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.

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.
- They also indicate that the common practice of adding interaction terms as excluded instruments may exacerbate the problem, even while reducing the standard error of the coefficient on the endogenous explanatory variable.
- The first assumption of IV method, thus revelance 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

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:

- 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 overidentification," that is, the number of instruments minus the number of endogenous regressors.

Application: Demand for Cigarettes

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

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

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).
- 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.

Instrumental Variable for multiple regression

IV for multiple regression(Key Concept 12.1)

- Our model is a multiple regression

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

- 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 $\hat{X}^*_{1i}$ 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 $(\hat{X}^*_{1i}, ..., \hat{X}^*_{ki}, W_1, X_2, ...W_r, 1)$ 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.