Instrumental Variables Regression

Outline

1. IV Regression: Why and What; Two Stage Least Squares
2. The General IV Regression Model
3. Checking Instrument Validity
   a) Weak and strong instruments
   b) Instrument exogeneity
4. Application: Demand for cigarettes
5. Examples: Where Do Instruments Come From?

IV Regression: Why?

Three important threats to internal validity are:

- Omitted variable bias from a variable that is correlated with X but is unobserved (so cannot be included in the regression) and for which there are inadequate control variables;
- Simultaneous causality bias (X causes Y, Y causes X);
- Errors-in-variables bias (X is measured with error)

All three problems result in $E(u|X) \neq 0$.

- Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ – using an *instrumental variable* (IV), Z.

The IV Estimator with a Single Regressor and a Single Instrument (SW Section 12.1)

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

- The goal is an estimate of the causal effect β_1. However, X is correlated with the error term, and we cannot solve the problem simply by including control variables.
- Instrumental variables (IV) regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an ***instrumental variable***, Z_i, which is correlated with X_i but uncorrelated with u_i.

Terminology: Endogeneity and Exogeneity

An ***endogenous*** variable is one that is correlated with u

An ***exogenous*** variable is one that is uncorrelated with u

In IV regression, we focus on the case that X is endogenous and there is an instrument, Z, which is exogenous.

Digression on terminology: "Endogenous" literally means "determined within the system." If X is jointly determined with Y, then a regression of Y on X is subject to simultaneous causality bias. But this definition of endogeneity is too narrow because IV regression can be used to address OV bias and errors-in-variable bias. Thus we use the broader definition of endogeneity above.

Two Conditions for a Valid Instrument

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

For an instrumental variable (an "***instrument***") Z to be valid, it must satisfy two conditions:

1. ***Instrument relevance***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i, u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later). How can you use Z_i to estimate β_1?

The IV estimator with one X and one Z (1 of 7)

Explanation #1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages – two regressions:

(1) Isolate the part of X that is uncorrelated with u by regressing X on Z using OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \quad (1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don' t know π_0 or π_1 but we have estimated them, so…

- Compute the predicted values of X_i, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

The IV estimator with one X and one Z (2 of 7)

(2) Replace X_i by $\hat{X}_i$ in the regression of interest: regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- **Because $\hat{X}_i$ is uncorrelated with u_i, the first least squares assumption holds for regression (2).** (This requires n to be large so that π_0 and π_1 are precisely estimated.)

- Thus, in large samples, β_1 can be estimated by OLS using regression (2)

- The resulting estimator is called the *Two Stage Least Squares* (*TSLS*) estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares: Summary

Suppose Z_i, satisfies the two conditions for a valid instrument:

1. ***Instrument relevance***: corr(Z_i, X_i) $\neq$ 0
2. ***Instrument exogeneity***: corr(Z_i, u_i) = 0

Two-stage least squares:

Stage 1: Regress X_i on Z_i (including an intercept), obtain the predicted values $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$ (including an intercept); the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

The IV estimator with one X and one Z (3 of 7)

Explanation #2: A direct algebraic derivation

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

Thus,

$$\begin{aligned}
\text{cov}(Y_i, Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0, Z_i) + \text{cov}(\beta_1 X_i, Z_i) + \text{cov}(u_i, Z_i) \\
&= 0 + \text{cov}(\beta_1 X_i, Z_i) + 0 \\
&= \beta_1 \text{cov}(X_i, Z_i)
\end{aligned}$$

where $\text{cov}(u_i, Z_i) = 0$ by instrument exogeneity; thus

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

The IV estimator with one X and one Z (4 of 7)

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

The IV estimator replaces these population covariances with sample covariances:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}},$$

where s_{YZ} and s_{XZ} are the sample covariances. This is the TSLS estimator – just a different derivation!

The IV estimator with one X and one Z (5 of 7)

Explanation #3: Derivation from the "reduced form"

The "reduced form" relates Y to Z and X to Z:

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

$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

where w_i is an error term. Because Z is exogenous, Z is uncorrelated with both v_i and w_i.

The idea: A unit change in Z_i results in a change in X_i of π_1 and a change in Y_i of γ_1. Because that change in X_i arises from the exogenous change in Z_i, that change in X_i is exogenous. Thus an exogenous change in X_i of π_1 units is associated with a change in Y_i of γ_1 units – so the effect on Y of an exogenous change in X is $\boldsymbol{\beta_1 = \gamma_1/\pi_1}$.

The IV estimator with one X and one Z (6 of 7)

The math:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$
$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

Solve the X equation for Z:

$$Z_i = -\pi_0/\pi_1 + (1/\pi_1)X_i - (1/\pi_1)v_i$$

Substitute this into the Y equation and collect terms:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 Z_i + w_i \\ &= \gamma_0 + \gamma_1[-\pi_0/\pi_1 + (1/\pi_1)X_i - (1/\pi_1)v_i] + w_i \\ &= [\gamma_0 - \pi_0\gamma_1/\pi_1] + (\gamma_1/\pi_1)X_i + [w_i - (\gamma_1/\pi_1)v_i] \\ &= \beta_0 + \beta_1 X_i + u_i, \end{aligned}$$

where $\beta_0 = \gamma_0 - \pi_0\gamma_1/\pi_1$, $\boldsymbol{\beta_1 = \gamma_1/\pi_1}$, and $u_i = w_i - (\gamma_1/\pi_1)v_i$.

The IV estimator with one X and one Z (7 of 7)

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

$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

yields

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

where

$$\beta_1 = \gamma_1 / \pi_1$$

Interpretation: An exogenous change in X_i of π_1 units is associated with a change in Y_i of γ_1 units – so the effect on Y of an exogenous unit change in X is $\beta_1 = \gamma_1/\pi_1$.

Example #1: Effect of Studying on Grades (1 of 6)

What is the effect on grades of studying for an additional hour per day?

Y = GPA

X = study time (hours per day)

Data: grades and study hours of college freshmen.

Would you expect the OLS estimator of β_1 (the effect on GPA of studying an extra hour per day) to be unbiased? Why or why not?

Example #1: Effect of Studying on Grades (2 of 6)

Stinebrickner, Ralph and Stinebrickner, Todd R. (2008) "The Causal Effect of Studying on Academic Performance," *The B.E. Journal of Economic Analysis & Policy*: Vol. 8: Iss. 1 (Frontiers), Article 14.

- n = 210 freshman at Berea College (Kentucky) in 2001
- Y = first-semester GPA
- X = average study hours per day (time use survey)
- Roommates were randomly assigned
- Z = 1 if roommate brought video game, = 0 otherwise

Do you think Z_i (whether a roommate brought a video game) is a valid instrument?

1. Is it relevant (correlated with X)?
2. Is it exogenous (uncorrelated with u)?

Example #1: Effect of Studying on Grades (3 of 6)

$X = \pi_0 + \pi_1 Z + v_i$

$Y = \gamma_0 + \gamma_1 Z + w_i$

$Y =$ *GPA* (4 point scale)

$X =$ *time spent studying (hours per day)*

$Z =$ *1 if roommate brought video game, = 0 otherwise*

Stinebrinckner and Stinebrinckner's findings

$$\hat{\pi}_1 = -.668$$

$$\hat{\gamma}_1 = -.241$$

$$\hat{\beta}_1^{IV} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-.241}{-.668} = 0.360$$

What are the units? Do these estimates make sense in a real-world way? (*Note*: They actually ran the regressions including additional regressors – more on this later.)

Consistency of the TSLS estimator

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

The sample covariances are consistent: $s_{YZ} \xrightarrow{p} \text{cov}(Y, Z)$

and $s_{XZ} \xrightarrow{p} \text{cov}(X, Z)$. Thus,

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y, Z)}{\text{cov}(X, Z)} = \beta_1$$

- The instrument relevance condition, cov(X, Z) $\neq$ 0, ensures that you don' t divide by zero.

Example #2: Supply and demand for butter (1 of 2)

IV regression was first developed to estimate demand elasticities for agricultural goods, for example, butter:

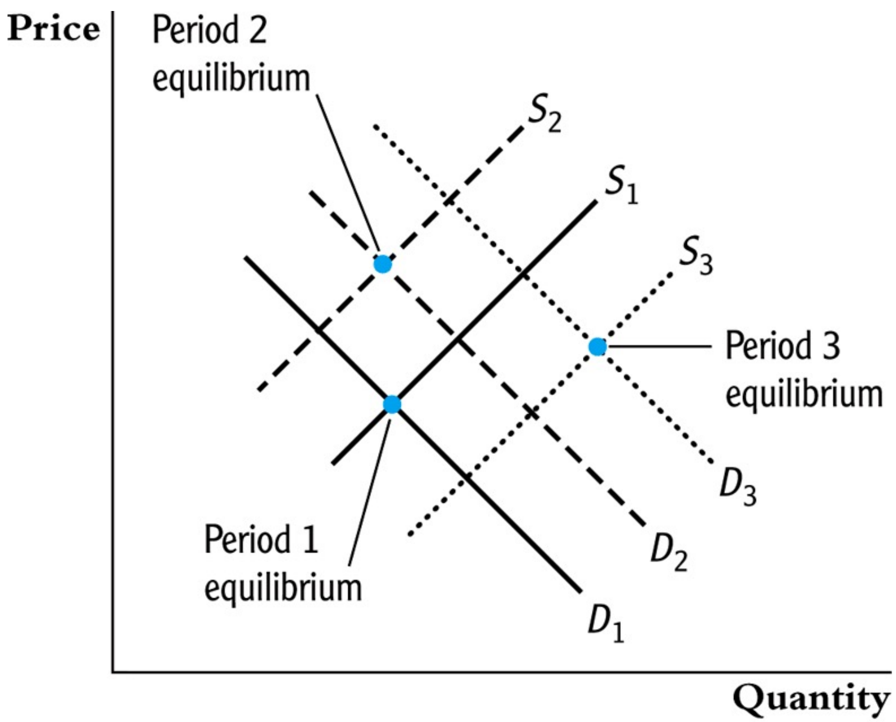
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- β_1 = price elasticity of demand for butter = percent change in quantity for a 1% change in price (recall log-log specification discussion)
- Data: observations on price and quantity of butter for different years

- The OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ suffers from simultaneous causality bias (*why*?)

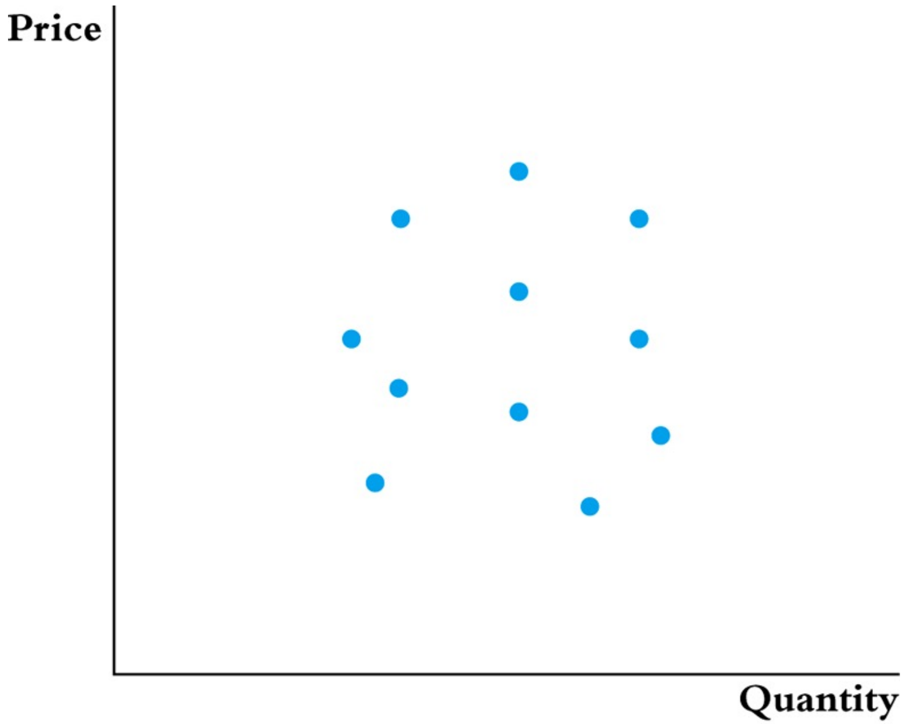
Example #2: Supply and demand for butter

Simultaneous causality bias in the OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ arises because price and quantity are determined by the interaction of demand *and* supply:



(a) Demand and supply in three time periods

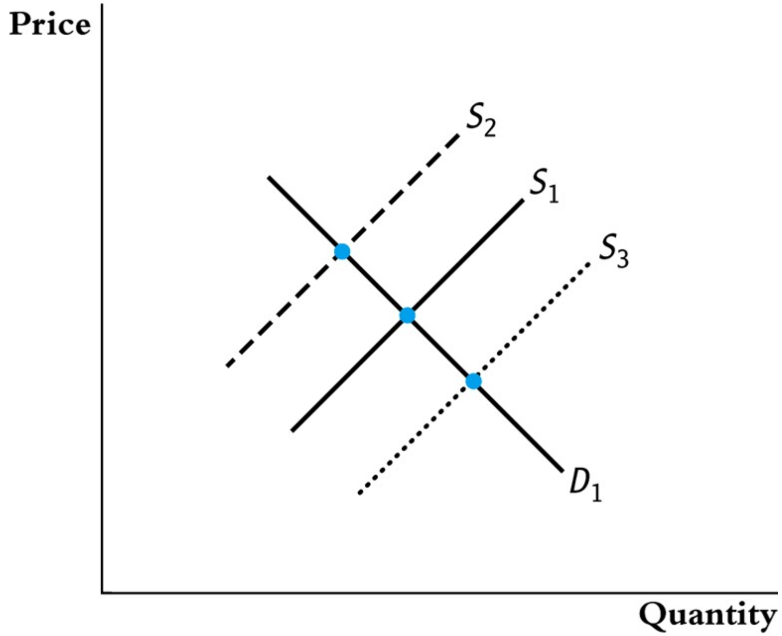
This interaction of demand and supply produces data like…



(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?

But···what would you get if only supply shifted?



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

- TSLS estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply.
- *Z* is a variable that shifts supply but not demand.

TSLS in the supply-demand example (1 of 2)

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

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

(1) Relevant? $\text{corr}(rain_i, \ln(P_i^{butter})) \neq 0$?

Plausibly: insufficient rainfall means less grazing means less butter means higher prices

(2) Exogenous? corr($rain_i$, u_i) = 0?

Plausibly: whether it rains in dairy-producing regions shouldn't affect demand for butter

TSLS in the supply-demand example (2 of 2)

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

Z_i = $rain_i$ = rainfall in dairy-producing regions.

Stage 1: regress $\ln(P_i^{butter})$ on *rain*, get $\ln(P_i^{butter})$

$\ln(P_i^{butter})$ isolates changes in log price that arise from supply (part of supply, at least)

Stage 2: regress $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$

The regression counterpart of using shifts in the supply curve to trace out the demand curve.

Example #3: Test scores and class size (1 of 2)

- The California test score/class size regressions still could have OV bias (e.g. parental involvement).
- In principle, this bias can be eliminated by IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:
  1. relevant: corr(Z_i, STR_i) $\neq$ 0
  2. exogenous: corr(Z_i, u_i) = 0

Example #3: Test scores and class size (2 of 2)

Here is a (hypothetical) instrument:

- some districts, randomly hit by an earthquake, "double up" classrooms:

$$Z_i = Quake_i = 1 \text{ if hit by quake, } = 0 \text{ otherwise}$$

- *Do the two conditions for a valid instrument hold*?
- The earthquake makes it *as if* the districts were in a random assignment experiment. Thus, the variation in *STR* arising from the earthquake is exogenous.
- The first stage of TSLS regresses *STR* against *Quake*, thereby isolating the part of *STR* that is exogenous (the part that is "as if" randomly assigned)

Inference using TSLS (1 of 5)

- In large samples, the sampling distribution of the TSLS estimator is normal
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. $\pm$ 1.96*SE*
- The idea behind the large-sample normal distribution of the TSLS estimator is that – like all the other estimators we have considered – it involves an average of mean zero i.i.d. random variables, to which we can apply the CLT.
- Here is the math (SW App. 12.3)···

Inference using TSLS (2 of 5)

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})} = \frac{\sum_{i=1}^{n} Y_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Substitute in $Y_i = \beta_0 + \beta_1 X_i + u_i$ and simplify:

$$\hat{\beta}_1^{TSLS} = \frac{\beta_1\sum_{i=1}^{n} X_i(Z_i - \bar{Z}) + \sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

so…

Inference using TSLS (3 of 5)

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}.$$

So $$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Multiply through by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Inference using TSLS (4 of 5)

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z})}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X, Z) \neq 0$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i \text{ is distributed } N(0, \text{var}[(Z - \mu_Z)u]) \quad \text{(CLT)}$$

so: $\hat{\beta}_1^{TSLS}$ is approx. distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$,

where $$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}.$$

where cov(X, Z) $\neq$ 0 because the instrument is relevant

Pearson

Inference using TSLS (5 of 5)

$$\hat{\beta}_1^{TSLS} \text{ is approx. distributed } N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}),$$

- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples
- This all assumes that the instruments are valid – we' ll discuss what happens if they aren' t valid shortly.
- ***Important note on standard errors***:

  – The OLS standard errors from the second stage regression aren't right – they don't take into account the estimation in the first stage ($\hat{X}_i$ is estimated).

  - Instead, use a single specialized command that computes the TSLS estimator and the correct *SE*s.
  - As usual, use heteroskedasticity-robust *SE*s

Example #4: Demand for Cigarettes (1 of 3)

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

Why is the OLS estimator of β_1 likely to be biased?

- Data set: Panel data on annual cigarette consumption and average prices paid (including tax), by state, for the 48 continental US states, 1985–1995.
- Proposed instrumental variable:
  - Z_i = general sales tax per pack in the state = $SalesTax_i$
  - Do you think this instrument is plausibly valid?

1. Relevant? $\text{corr}(SalesTax_i, \ln(P_i^{cigarettes})) \neq 0$?
2. Exogenous? $\text{corr}(SalesTax_i, u_i) = 0$?

Example #4: Demand for Cigarettes (2 of 3)

For now, use data from 1995 only.

First stage OLS regression:

$$\widehat{\ln(P_i^{cigarettes})} = 4.63 + .031 SalesTax_i, \; n = 48$$

Second stage OLS regression:

$$\widehat{\ln(Q_i^{cigarettes})} = 9.72 - 1.08 \; \widehat{\ln(P_i^{cigarettes})}, \; n = 48$$

Combined TSLS regression with correct, heteroskedasticity-robust standard errors:

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.53)}{9.72} - \underset{(0.32)}{1.08}, \ln(P_i^{cigarettes}) \; n = 48$$

2SLS in R(1 of 4)

```
# load the data set and get an overview
library(AER)
data("CigarettesSW")
summary(CigarettesSW)
#>     state      year        cpi          population          packs
#>  AL     : 2   1985:48   Min.   :1.076   Min.   :  478447   Min.   : 49.27
#>  AR     : 2   1995:48   1st Qu.:1.076   1st Qu.: 1622606   1st Qu.: 92.45
#>  AZ     : 2             Median :1.300   Median : 3697472   Median :110.16
#>  CA     : 2             Mean   :1.300   Mean   : 5168866   Mean   :109.18
#>  CO     : 2             3rd Qu.:1.524   3rd Qu.: 5901500   3rd Qu.:123.52
#>  CT     : 2             Max.   :1.524   Max.   :31493524   Max.   :197.99
#>  (Other):84
#>     income               tax            price             taxs
#>  Min.   :  6887097   Min.   :18.00   Min.   : 84.97   Min.   : 21.27
#>  1st Qu.: 25520384   1st Qu.:31.00   1st Qu.:102.71   1st Qu.: 34.77
#>  Median : 61661644   Median :37.00   Median :137.72   Median : 41.05
#>  Mean   : 99878736   Mean   :42.68   Mean   :143.45   Mean   : 48.33
#>  3rd Qu.:127313964   3rd Qu.:50.88   3rd Qu.:176.15   3rd Qu.: 59.48
#>  Max.   :771470144   Max.   :99.00   Max.   :240.85   Max.   :112.63
#>
```

2SLS R(2 of 4)

```
# compute real per capita prices
CigarettesSW$rprice <- with(CigarettesSW, price / cpi)

#  compute the sales tax
CigarettesSW$salestax <- with(CigarettesSW, (taxs - tax) / cpi)

# check the correlation between sales tax and price
cor(CigarettesSW$salestax, CigarettesSW$price)
#> [1] 0.6141228

# generate a subset for the year 1995
c1995 <- subset(CigarettesSW, year == "1995")

# perform the first stage regression
cig_s1 <- lm(log(rprice) ~ salestax, data = c1995)

coeftest(cig_s1, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error  t value  Pr(>|t|)
#> (Intercept) 4.6165463  0.0289177 159.6444 < 2.2e-16 ***
#> salestax    0.0307289  0.0048354   6.3549 8.489e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2SLS R(3 of 4)

```
# store the predicted values
lcigp_pred <- cig_s1$fitted.values

# run the stage 2 regression
cig_s2 <- lm(log(c1995$packs) ~ lcigp_pred)
coeftest(cig_s2, vcov = vcovHC)
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)  9.71988    1.70304  5.7074 7.932e-07 ***
#> lcigp_pred  -1.08359    0.35563 -3.0469  0.003822 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\ln(Q_i^{cigarettes}) = \underset{(1.53)}{9.72} - \underset{(0.31)}{1.08} \ln(P_i^{cigarettes}), \; n = 48$$

2SLS R(4 of 4)

```
# perform TSLS using 'ivreg()'
cig_ivreg <- ivreg(log(packs) ~ log(rprice) | salestax, data = c1995)

coeftest(cig_ivreg, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)  9.71988    1.52832  6.3598 8.346e-08 ***
#> log(rprice) -1.08359    0.31892 -3.3977  0.001411 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\ln(Q_i^{cigarettes}) = \underset{(1.53)}{9.72} - \underset{(0.31)}{1.08} \ln(P_i^{cigarettes}), \; n = 48$$

Summary of IV Regression with a Single X and Z

- A valid instrument Z must satisfy two conditions:
  1. *relevance*: corr(Z_i, X_i) $\neq$ 0
  2. *exogeneity*: corr(Z_i, u_i) = 0
- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$
- The key idea is that the first stage isolates part of the variation in X that is uncorrelated with u
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual

The General IV Regression Model (SW Section 12.2)

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors ($X_1,\cdots, X_k$)
  - multiple included exogenous variables ($W_1,\cdots,W_r$) or control variables
  - multiple instrumental variables ($Z_1,\ldots,Z_m$). Having more (relevant) instruments can produce a smaller variance of TSLS: the R^2 of the first stage increases, so you have more variation in $\hat{X}$.
- ***New terminology***: identification & overidentification

Identification (1 of 2)

- In general, a parameter is said to be ***identified*** if different values of the parameter produce different distributions of the data.
- In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (*m*) and the number of endogenous regressors (*k*)
- Intuitively, if there are fewer instruments than endogenous regressors, we can' t estimate $\beta_1, \cdots, \beta_k$
  - For example, suppose $k = 1$ but $m = 0$ (no instruments)!

Identification (2 of 2)

The coefficients $\beta_1, \cdots, \beta_k$ are said to be:

- ***exactly identified*** if $m = k$.

  There are just enough instruments to estimate $\beta_1, \cdots, \beta_k$.

- ***overidentified*** if $m > k$.

  There are more than enough instruments to estimate $\beta_1, \cdots, \beta_k$. If so, you can test whether the instruments are valid (a test of the "overidentifying restrictions") – we'll return to this later

- ***underidentified*** if $m < k$.

  There are too few instruments to estimate $\beta_1, \cdots, \beta_k$. If so, you need to get more instruments!

The General IV Regression Model: Summary of Jargon

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

- Y_i is the ***dependent variable***
- $X_{1i}, \cdots, X_{ki}$ are the ***endogenous regressors*** (potentially correlated with u_i)
- $W_{1i}, \cdots, W_{ri}$ are the ***included exogenous regressors*** (uncorrelated with u_i) or ***control variables*** (included so that Z_i is uncorrelated with u_i, once the W s are included)
- $\beta_0, \beta_1, \cdots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i}, \cdots, Z_{mi}$ are the m ***instrumental variables*** (the ***excluded exogenous variables***)
- The coefficients are ***overidentified*** if $m > k$; ***exactly identified*** if $m = k$; and ***underidentified*** if $m < k$.

TSLS with a Single Endogenous Regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

- m instruments: $Z_{1i}, \cdots, Z_m$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1, \cdots, W_r, Z_1, \cdots, Z_m$, and an intercept, by OLS
    - Compute predicted values $\hat{X}_{1i}, i = 1, \ldots, n$
- Second stage
    - Regress Y on $\hat{X}_{1i}, W_1, \ldots, W_r$, and an intercept, by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SE*s are wrong
- To get correct *SE*s, do this in a single step in your regression software

Pearson

Example #4: Demand for cigarettes (3 of 3)

Suppose income is exogenous (this is plausible – *why?*), and we also want to estimate the income elasticity:

$$\ln(\ln(Q_i^{cigarettes})) = \beta_0 + \beta_1 \ln(\ln(P_i^{cigarettes})) + \beta_2 \ln(Income_i) + u_i$$

We actually have two instruments:

Z_{1i} = general sales tax$_i$

Z_{2i} = cigarette-specific tax$_i$

- Endogenous variable: $\ln(\ln(P_i^{cigarettes}))$ ("one X")
- Included exogenous variable: $\ln(Income_i)$ ("one W")
- Instruments (excluded endogenous variables): general sales tax, cigarette-specific tax ("two Zs")
- *Is β_1 over–, under–, or exactly identified?*

Example: Cigarette demand, one instrument

```
# add rincome to the dataset
CigarettesSW$rincome <- with(CigarettesSW, income / population / cpi)

c1995 <- subset(CigarettesSW, year == "1995")

# estimate the model
cig_ivreg2 <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
                    salestax, data = c1995)

coeftest(cig_ivreg2, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)   9.43066    1.25939  7.4883 1.935e-09 ***
#> log(rprice)  -1.14338    0.37230 -3.0711  0.003611 **
#> log(rincome)  0.21452    0.31175  0.6881  0.494917
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Cigarette demand, two instruments

(1 of 2)

```
# add cigtax to the data set
CigarettesSW$cigtax <- with(CigarettesSW, tax/cpi)

c1995 <- subset(CigarettesSW, year == "1995")

# estimate the model
cig_ivreg3 <- ivreg(log(packs) ~ log(rprice) + log(rincome) |  log(rincome) + salestax
  + cigtax, data = c1995)

coeftest(cig_ivreg3, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>               Estimate Std. Error t value  Pr(>|t|)
#> (Intercept)    9.89496    0.95922 10.3157 1.947e-13 ***
#> log(rprice)   -1.27742    0.24961 -5.1177 6.211e-06 ***
#> log(rincome)   0.28040    0.25389  1.1044    0.2753
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pearson

Example: Cigarette demand, two instruments (2 of 2)

TSLS estimates, Z = sales tax (m = 1)

$$\ln(Q_i^{cigarettes}) = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14} \ln(P_i^{cigarettes}) + \underset{(0.31)}{0.21} \ln(Income_i)$$

TSLS estimates, Z = sales tax & cig-only tax (m = 2)

$$\ln(Q_i^{cigarettes}) = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28} \ln(P_i^{cigarettes}) + \underset{(0.25)}{0.28} \ln(Income_i)$$

- Smaller *SEs* for $m = 2$. Using 2 instruments gives more information – more "as-if random variation."
- Low income elasticity (not a luxury good); income elasticity not statistically significantly different from 0
- Surprisingly high price elasticity

The General Instrument Validity Assumptions

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

(1) ***Instrument exogeneity***: corr(Z_{1i}, u_i) = 0,⋯, corr(Z_{mi}, u_i) = 0

(2) ***Instrument relevance***: *General case, multiple X' s*

Suppose the second stage regression could be run using the predicted values from the *population* first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression.

- *Special case of one X*: the general assumption is equivalent to (a) at least one instrument must enter the population counterpart of the first stage regression, and (b) the *W*' s are not perfectly multicollinear.

The IV Regression Assumptions

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

1. $E(u_i | W_{1i}, \cdots, W_{ri}) = 0$
   - #1 says "the exogenous regressors are exogenous."
2. $(Y_i, X_{1i}, \cdots, X_{ki}, W_{1i}, \cdots, W_{ri}, Z_{1i}, \cdots, Z_{mi})$ are i.i.d.
   - #2 is not new
3. The X's, W's, Z's, and Y have nonzero, finite 4th moments
   - #3 is not new
4. The instruments $(Z_{1i}, \cdots, Z_{mi})$ are valid.
   - We have discussed this

- Under 1–4, TSLS and its t-statistic are normally distributed
- The critical requirement is that the instruments be valid

W's as control variables (1 of 2)

- In many cases, the purpose of including the W's is to control for omitted factors, so that once the W's are included, Z is uncorrelated with u. If so, W's don't need to be exogenous; instead, the W's need to be effective control variables in the sense discussed in Chapter 7 – except now with a focus on producing an exogenous instrument.
- Technically, the condition for W's being effective control variables is that the conditional mean of u_i does not depend on Z_i, given W_i:

$$E(u_i | W_i, Z_i) = E(u_i | W_i)$$

W's as control variables (2 of 2)

- Thus an alternative to IV regression assumption #1 is that conditional mean independence holds:

$$E(u_i|W_i, Z_i) = E(u_i|W_i)$$

  This is the IV version of the conditional mean independence assumption in Chapter 7.

- *Here is the key idea*: in many applications you need to include control variables (W's) so that Z is plausibly exogenous (uncorrelated with u).

- For the math, see SW Appendix 12.6. For an example, see…

Example #1: Effect of Studying on Grades (4 of 6)

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

Y = first-semester GPA

X = average study hours per day

Z = 1 if roommate brought video game, = 0 otherwise

Roommates were randomly assigned

Can you think of a reason that Z might be correlated with u – even though it is randomly assigned? What else enters the error term – what are other determinants of grades, beyond time spent studying?

Example #1: Effect of Studying on Grades (5 of 6)

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

Why might Z be correlated with u?

- Here's a hypothetical possibility: the student's sex. Suppose:
  - Roommates are randomly assigned – except always men with men and women with women.
  - Women get better grades than men, holding constant hour spent studying
  - Men are more likely to bring a video game than women
  - Then corr(Z_i, u_i) < 0 (males are more likely to have a [male] roommate who brings a video game – but males also tend to have lower grades, holding constant the amount of studying).
- Because corr(Z_i, u_i) < 0, the IV (roommate brings video game) isn't valid.
  - This is the IV version of OV bias.
  - The solution to OV bias is to control for (or include) the OV – in this case, sex.

Example #1: Effect of Studying on Grades (6 of 6)

- This logic leads you to include W = student's sex as a control variable in the IV regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

- The TSLS estimate reported above is from a regression that included gender as a W variable – along with other variables such as individual i's major.
- The conditional mean independence condition for an exogenous instrument is, $E(u_i|Z_i, W_i) = E(u_i|W_i)$.
  - In words: among men (conditional on W = male), roommates are randomly assigned, so whether your roommate brings a video game is random. Same thing among women (conditional on W = female).
  - The instrument is not exogenous if W isn't included in the regression.
  - But when W is included, the conditional mean independence condition $E(u_i|Z_i, W_i) = E(u_i|W_i)$ holds, and the instrument is valid.

Checking Instrument Validity (SW Section 12.3)

Recall the two requirements for valid instruments:

1. *Relevance* (special case of one X)
   At least one instrument must enter the population counterpart of the first stage regression.
2. *Exogeneity*
   All the instruments must be uncorrelated with the error term:
   $\text{corr}(Z_{1i}, u_i) = 0, \cdots, \text{corr}(Z_{mi}, u_i) = 0$

What happens if one of these requirements isn't satisfied? How can you check? What do you do?

If you have multiple instruments, which should you use?

Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

First stage regression:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+k} W_{ki} + u_i$$

- The instruments are relevant if at least one of $\pi_1, \cdots, \pi_m$ are nonzero.
- The instruments are said to be ***weak*** if all the $\pi_1, \cdots, \pi_m$ are either zero or nearly zero.
- ***Weak instruments*** explain very little of the variation in X, beyond that explained by the W s

What are the consequences of weak instruments?

If instruments are weak, the sampling distribution of TSLS and its *t*-statistic are not (at all) normal, even with *n* large.

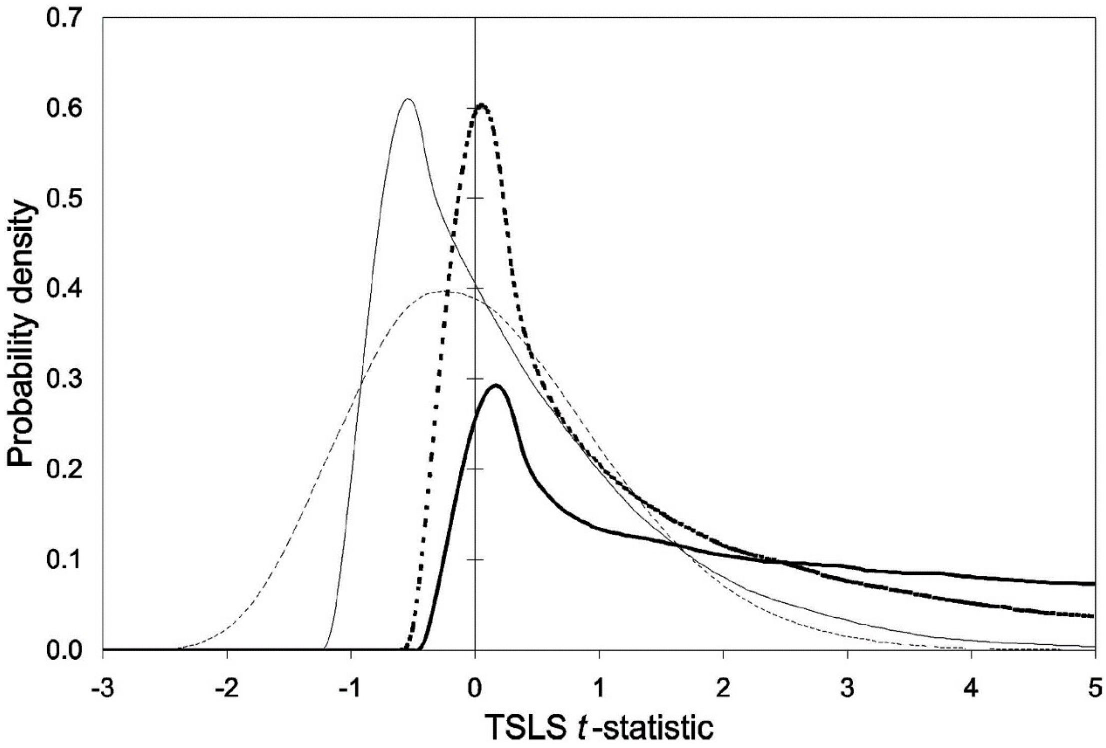
Consider the simplest case of 1 X, 1 Z, no control variables:

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

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

- The IV estimator is $\hat{\beta}_1^{TSLS} = \dfrac{s_{YZ}}{s_{XZ}}$
  - If cov(X, Z) is zero or small, then s_{XZ} will be small: With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of $\hat{\beta}_1^{TSLS}$ (and its t-statistic) is not well approximated by its large-n normal approximation...

An example: The sampling distribution of the TSLS *t*-statistic with weak instruments



Dark line = irrelevant instruments

Dashed light line = strong instruments

Why does our trusty normal approximation fail us?

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- If $\text{cov}(X, Z)$ is small, small changes in s_{XZ} (from one sample to the next) can induce big changes in $\hat{\beta}_1^{TSLS}$
  - Suppose in one sample you calculate $s_{XZ} = .00001$…
- Thus the large-n normal approximation is a poor approximation to the sampling distribution of $\hat{\beta}_1^{TSLS}$
- A better approximation is that $\hat{\beta}_1^{TSLS}$ is distributed as the *ratio* of two correlated normal random variables (see SW App. 12.4)
  - If instruments are weak, the usual methods of inference are unreliable – potentially very unreliable.

Measuring the Strength of Instruments in Practice: The First-Stage F-statistic

- The first stage regression (one X):
- Regress X on $Z_1,..,Z_m, W_1,\cdots,W_k$.
- Totally irrelevant instruments ↔ *all* the coefficients on $Z_1,\cdots,Z_m$ are zero.
- The ***first-stage F-statistic*** tests the hypothesis that $Z_1,\cdots,Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

Checking for Weak Instruments with a Single X (1 of 2)

- Compute the first-stage F-statistic.

  Rule-of-thumb: If the first stage F-statistic is less than 10, then the set of instruments is weak.

- If so, the TSLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.

Checking for Weak Instruments with a Single X (2 of 2)

- Why compare the first-stage F to 10?
- Simply rejecting the null hypothesis that the coefficients on the Z s are zero isn't enough – you need substantial predictive content for the normal approximation to be a good one.
- Comparing the first-stage F to 10 tests for whether the bias of TSLS, relative to OLS, is less than 10%. If F is smaller than 10, the relative bias exceeds 10%—that is, TSLS can have substantial bias (see SW App. 12.5).

What to do if you have weak instruments

- Get better instruments (often easier said than done!)
- If you have many instruments, some are probably weaker than others and it's a good idea to drop the weaker ones (dropping an irrelevant instrument will increase the first-stage F)
- If you only have a few instruments, and all are weak, then you need to do some IV analysis other than TSLS…
  - Separate the problem of estimation of β_1 and construction of confidence intervals
  - This seems odd, but if TSLS isn't normally distributed, it makes sense (right?)

Confidence Intervals with Weak Instruments (1 of 2)

- With weak instruments, TSLS confidence intervals are not valid – but some other confidence intervals *are*. Here are two ways to compute confidence intervals that are valid in large samples, even if instruments are weak:

1. The Anderson-Rubin confidence interval
   - The Anderson-Rubin confidence interval is based on the Anderson-Rubin test statistic testing $\beta_1 = \beta_{1,0}$:
     - Compute $= Y_i - \beta_{1,0}X_i$
     - Regress on $W_{1i}, \cdots, W_{ri}, Z_{1i}, \cdots, Z_{mi}$
     - The AR test is the F-statistic on $Z_{1i}, \cdots, Z_{mi}$
   - Now invert this test: the 95% AR confidence interval is the set of β_1 *not* rejected at the 5% level by the AR test.
   - Computation: a pain by hand! use specialized software.

Confidence Intervals with Weak Instruments (2 of 2)

2. Moreira's Conditional Likelihood Ratio confidence interval
    - The Conditional Likelihood Ratio (CLR) confidence interval is based on inverting Moreira's Conditional Likelihood Ratio test. Computing this test, its critical value, and the CLR confidence interval requires specialized software.
    - The CLR confidence interval tends to be tighter than the Anderson-Rubin confidence interval, especially when there are many instruments.
    - If your software produces the CLR confidence interval, this is the one to use.

Pearson

Weak Instruments and Heteroskedasticity

The foregoing discussion applies to the homoskedasticity case. In practice, you would want to use robust SEs, either heteroskedasticity-robust or, in panel data, clustered SEs.

- If you have 1 X and 1 Z:
  - Assess instrument strength using the robust first-stage F, which you can compare to 10
  - Compute weak-instrument confidence intervals by the Anderson-Rubin method, using robust SEs in the regression of $Y_i - \beta_{1,0}X_i$ on $W_{1i},\ldots, W_{ri}$, $Z_{1i},\ldots, Z_{mi}$
- If you have more than one Z, then the methods for weak-instrument robust inference go beyond the scope of this book. A reasonable compromise – better than ignoring the weak instrument problem – is to use homoskedasticity-only SEs for the first stage F and the CLR (if available) for confidence intervals for β_1

Estimation with Weak Instruments

There are no unbiased estimators if instruments are weak or irrelevant. However, some estimators have a distribution more centered around β_1 than TSLS.

- One such estimator is the limited information maximum likelihood estimator (LIML)
- The LIML estimator
  - can be derived as a maximum likelihood estimator
  - is the value of β_1 that minimizes the p-value of the AR test(!)
- For more discussion about estimators, tests, and confidence intervals when you have weak instruments, see SW, App. 12.5

Checking Assumption #2: Instrument Exogeneity

- Instrument exogeneity: ***All*** the instruments are uncorrelated with the error term: corr(Z_{1i}, u_i) = 0,⋯, corr(Z_{mi}, u_i) = 0

- If the instruments are correlated with the error term, the first stage of TSLS cannot isolate a component of X that is uncorrelated with the error term, so $\hat{X}$ is correlated with u and TSLS is inconsistent.

- If there are more instruments than endogenous regressors, it is possible to test – *partially* – for instrument exogeneity.

Testing Overidentifying Restrictions

Consider the simplest case:

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

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}
- 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 J-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if $\#Z$'s > $\#X$'s (overidentified).

The J-test of Overidentifying Restrictions

Suppose # instruments = m > # X' s = k (overidentified)

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

The J-test is the Anderson-Rubin test, using the TSLS estimator instead of the hypothesized value $\beta_{1,0}$. The recipe:

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$, using the *actual* X's (not the $\hat{X}$'s used to estimate the second stage)
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regress against $Z_{1i}, \cdots, Z_{mi}, W_{1i}, \cdots, W_{ri}$
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i}, \cdots, Z_{mi}$ are all zero;
5. The ***J-statistic*** is $J = mF$

Pearson

The J-test of Overidentifying Restrictions (2 of 2)

$J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\cdots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\cdots,Z_{mi},\ W_{1i},\cdots,W_{ri}$.

Distribution of the J-statistic

- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m–k$ degrees of freedom
- If $m = k$, $J = 0$ (*does this make sense?*)
- If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.

Checking Instrument Validity: Summary (1 of 2)

This summary considers the case of a single *X*. The two requirements for valid instruments are:

1. ***Relevance***

- At least one instrument must enter the population counterpart of the first stage regression.
- If instruments are weak, then the TSLS estimator is biased and the and *t*-statistic has a non-normal distribution
- To check for weak instruments with a single included endogenous regressor, check the first-stage *F*
  - If $F > 10$, instruments are strong – use TSLS
  - If $F < 10$, weak instruments – take some action.

Checking Instrument Validity: Summary (2 of 2)

2. ***Exogeneity***
   - ***All*** the instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i}, u_i) = 0, \cdots, \text{corr}(Z_{mi}, u_i) = 0$
   - We can partially test for exogeneity: if $m > 1$, we can test the null hypothesis that all the instruments are exogenous, against the alternative that as many as $m - 1$ are endogenous (correlated with u)
   - The test is the J-test, which is constructed using the TSLS residuals.
   - If the J-test rejects, then at least some of your instruments are endogenous – so you must make a difficult decision and jettison some (or all) of your instruments.

Application to the Demand for Cigarettes (SW Section 12.4)

Why are we interested in knowing the elasticity of demand for cigarettes?

- Theory of optimal taxation. The optimal tax rate is inversely related to the price elasticity: the greater the elasticity, the less quantity is affected by a given percentage tax, so the smaller is the change in consumption and deadweight loss.
- Externalities of smoking – role for government intervention to discourage smoking
  - health effects of second-hand smoke? (non-monetary)
  - monetary externalities

Panel data set

- Annual cigarette consumption, average prices paid by end consumer (including tax), personal income, and tax rates (cigarette-specific and general statewide sales tax rates)
- 48 continental US states, 1985–1995

Estimation strategy

- We need to use IV estimation methods to handle the simultaneous causality bias that arises from the interaction of supply and demand.
- State binary indicators = W variables (control variables) which control for unobserved state-level characteristics that affect the demand for cigarettes and the tax rate, as long as those characteristics don't vary over time.

Fixed-effects model of cigarette demand

$$\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + u_{it}$$

- $i = 1, \cdots, 48$, $t = 1985, 1986, \cdots, 1995$
- $\text{corr}(\ln(P_{it}^{cigarettes}), u_{it})$ is plausibly nonzero because of supply/demand interactions
- α_i reflects unobserved omitted factors that vary across states but not over time, e.g. attitude towards smoking
- Estimation strategy:
  - Use panel data regression methods to eliminate α_i
  - Use TSLS to handle simultaneous causality bias
  - Use $T = 2$ with 1985 – 1995 changes ("changes" method) – look at long-term response, not short-term dynamics (short- v. long-run elasticities)

The "changes" method (when T=2)

- One way to model long-term effects is to consider 10-year changes, between 1985 and 1995
- Rewrite the regression in "changes" form:

$$\ln(Q_{i1995}^{cigarettes}) - \ln(Q_{i1985}^{cigarettes}) = \beta_1[\ln(P_{i1995}^{cigarettes}) - \ln(P_{i1985}^{cigarettes})]$$
$$+ \beta_2[\ln(Income_{i1995}) - \ln(Income_{i1985})] + (u_{i1995} - u_{i1985})$$

- Create "10-year change" variables, for example:
- 10-year change in log price = $\ln(P_{i1995}) - \ln(P_{i1985})$
- Then estimate the demand elasticity by TSLS using 10-year changes in the instrumental variables
- This is equivalent to using the original data and including the state binary indicators ("W" variables) in the regression

R: Cigarette demand

First create "10-year change" variables

10-year change in log price

$= \ln(P_{it}) - \ln(P_{it-10}) = \ln(P_{it}/P_{it-10})$

```
# subset data for year 1985
c1985 <- subset(CigarettesSW, year == "1985")

# define differences in variables
packsdiff <- log(c1995$packs) - log(c1985$packs)

pricediff <- log(c1995$price/c1995$cpi) - log(c1985$price/c1985$cpi)

incomediff <- log(c1995$income/c1995$population/c1995$cpi) -
log(c1985$income/c1985$population/c1985$cpi)

salestaxdiff <- (c1995$taxs - c1995$tax)/c1995$cpi - (c1985$taxs - c1985$tax)/c1985$cpi

cigtaxdiff <- c1995$tax/c1995$cpi - c1985$tax/c1985$cpi
```

Use TSLS to estimate the demand elasticity by using the "10-year changes" specification

```
# estimate the three models
cig_ivreg_diff1 <- ivreg(packsdiff ~ pricediff + incomediff | incomediff +  salestaxdiff)

cig_ivreg_diff2 <- ivreg(packsdiff ~ pricediff + incomediff | incomediff + cigtaxdiff)

cig_ivreg_diff3 <- ivreg(packsdiff ~ pricediff + incomediff | incomediff + salestaxdiff +
  cigtaxdiff)


# robust coefficient summary for 1.
coeftest(cig_ivreg_diff1, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept) -0.117962   0.068217 -1.7292   0.09062 .
#> pricediff   -0.938014   0.207502 -4.5205 4.454e-05 ***
#> incomediff   0.525970   0.339494  1.5493   0.12832
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use TSLS to estimate the demand elasticity by using the "10-year changes" specification

```
# robust coefficient summary for 2.
coeftest(cig_ivreg_diff2, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept) -0.017049   0.067217 -0.2536    0.8009
#> pricediff   -1.342515   0.228661 -5.8712 4.848e-07 ***
#> incomediff   0.428146   0.298718  1.4333    0.1587
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# robust coefficient summary for 3.
coeftest(cig_ivreg_diff3, vcov = vcovHC, type = "HC1")
#>
#> t test of coefficients:
#>
#>              Estimate Std. Error t value  Pr(>|t|)
#> (Intercept) -0.052003   0.062488 -0.8322    0.4097
#> pricediff   -1.202403   0.196943 -6.1053 2.178e-07 ***
#> incomediff   0.462030   0.309341  1.4936    0.1423
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check instrument relevance: compute first-stage F

```
# first-stage regressions
mod_relevance1 <- lm(pricediff ~ salestaxdiff + incomediff)
mod_relevance2 <- lm(pricediff ~ cigtaxdiff + incomediff)
mod_relevance3 <- lm(pricediff ~ incomediff + salestaxdiff + cigtaxdiff)

# check instrument relevance for model (1)
linearHypothesis(mod_relevance1,
                 "salestaxdiff = 0",
                 vcov = vcovHC, type = "HC1")
#> Linear hypothesis test
#>
#> Hypothesis:
#> salestaxdiff = 0
#>
#> Model 1: restricted model
#> Model 2: pricediff ~ salestaxdiff + incomediff
#>
#> Note: Coefficient covariance matrix supplied.
#>
#>   Res.Df Df      F    Pr(>F)
#> 1     46
#> 2     45  1 28.445 3.009e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check instrument relevance: compute first-stage F

```
# check instrument relevance for model (2)
linearHypothesis(mod_relevance2,
                 "cigtaxdiff = 0",
                 vcov = vcovHC, type = "HC1")
#> Linear hypothesis test
#>
#> Hypothesis:
#> cigtaxdiff = 0
#>
#> Model 1: restricted model
#> Model 2: pricediff ~ cigtaxdiff + incomediff
#>
#> Note: Coefficient covariance matrix supplied.
#>
#>   Res.Df Df      F   Pr(>F)
#> 1     46
#> 2     45  1 98.034 7.09e-13 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check instrument relevance: compute first-stage *F*

```
# check instrument relevance for model (3)
linearHypothesis(mod_relevance3,
                 c("salestaxdiff = 0", "cigtaxdiff = 0"),
                 vcov = vcovHC, type = "HC1")
#> Linear hypothesis test
#>
#> Hypothesis:
#> salestaxdiff = 0
#> cigtaxdiff = 0
#>
#> Model 1: restricted model
#> Model 2: pricediff ~ incomediff + salestaxdiff + cigtaxdiff
#>
#> Note: Coefficient covariance matrix supplied.
#>
#>   Res.Df Df      F    Pr(>F)
#> 1     46
#> 2     44  2 76.916 4.339e-15 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test the overidentifying restrictions (1 of 2)

```
# compute the J-statistic
cig_iv_OR <- lm(residuals(cig_ivreg_diff3) ~ incomediff + salestaxdiff + cigtaxdiff)

cig_OR_test <- linearHypothesis(cig_iv_OR,
                               c("salestaxdiff = 0", "cigtaxdiff = 0"),
                               test = "Chisq")
cig_OR_test
#> Linear hypothesis test
#>
#> Hypothesis:
#> salestaxdiff = 0
#> cigtaxdiff = 0
#>
#> Model 1: restricted model
#> Model 2: residuals(cig_ivreg_diff3) ~ incomediff + salestaxdiff + cigtaxdiff
#>
#>   Res.Df     RSS Df Sum of Sq Chisq Pr(>Chisq)
#> 1     46 0.37472
#> 2     44 0.33695  2  0.037769 4.932    0.08492 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test the overidentifying restrictions (2 of 2)

The correct degrees of freedom for the *J*-statistic is $m–k$:

- $J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\cdots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\cdots,Z_{mi}$, $W_{1i},\cdots,W_{mi}$.
- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m–k$ degrees of freedom
- Here, $J = 4.93$, distributed chi-squared with d.f. = 1; the 5% critical value is 3.84, so reject at 5% sig. level.

```
# compute correct p-value for J-statistic
pchisq(cig_OR_test[2, 5], df = 1, lower.tail = FALSE)
#> [1] 0.02636406
```

Tabular summary of these results:

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

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

How should we interpret the *J*-test rejection?

- *J*-test rejects the null hypothesis that both the instruments are exogenous
- This means that either *rtaxso* is endogenous, or *rtax* is endogenous, or both!
- The *J*-test doesn’ t tell us which! *You must exercise judgment*…
- Why might *rtax* (cig-only tax) be endogenous?
  - Political forces: history of smoking or lots of smokers ⇒ political pressure for low cigarette taxes
  - If so, cig-only tax is endogenous
- This reasoning doesn’ t apply to general sales tax
- → use just one instrument, the general sales tax

The Demand for Cigarettes: Summary of Empirical Results

- Use the estimated elasticity based on TSLS with the general sales tax as the only instrument:

$$\text{Elasticity} = -.94,\ SE = .21$$

- This elasticity is surprisingly large (not inelastic) – a 1% increase in prices reduces cigarette sales by nearly 1%. This is much more elastic than conventional wisdom in the health economics literature.
- This is a long-run (ten-year change) elasticity. *What would you expect a short-run (one-year change) elasticity to be – more or less elastic?*

Assess the Validity of the Study

Remaining threats to internal validity?

1. Omitted variable bias?
   - *The fixed effects estimator controls for unobserved factors that vary across states but not over time*
2. Functional form mis-specification? (*could check this*)
3. Remaining simultaneous causality bias?
   - *Not if the general sales tax a valid instrument, once state fixed effects are included!*
4. Errors-in-variables bias?
5. Selection bias? (*no, we have all the states*)
6. An additional threat to internal validity of IV regression studies is whether the instrument is (1) relevant and (2) exogenous. *How significant are these threats in the cigarette elasticity application?*

Assess the Validity of the Study (2 of 2)

External validity?

- We have estimated a long-run elasticity – can it be generalized to a short-run elasticity? Why or why not?
- Suppose we want to use the estimated elasticity of –0.94 to guide policy today. Here are two changes since the period covered by the data (1985–95) – do these changes pose a threat to external validity (generalization from 1985–95 to today)?
  - Levels of smoking today are lower than in 1985–1995
  - Cultural attitudes toward smoking have changed against smoking since 1985–95.

Where Do Valid Instruments Come From? (SW Section 12.5)

General comments

The hard part of IV analysis is finding valid instruments

- Method #1: "variables in another equation" (e.g. supply shifters that do not affect demand)
- Method #2: look for exogenous variation (Z) that is "as if" randomly assigned (does not directly affect Y) but affects X.
- These two methods are different ways to think about the same issues – see the link…
  - Rainfall shifts the supply curve for butter but not the demand curve; rainfall is "as if" randomly assigned
  - Sales tax shifts the supply curve for cigarettes but not the demand curve; sales taxes are "as if" randomly assigned

Example: Cardiac Catheterization (1 of 3)

McClellan, Mark, Barbara J. McNeil, and Joseph P. Newhouse (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association*, vol. 272, no. 11, pp. 859 – 866.

Does cardiac catheterization improve longevity of heart attack patients?

Y_i = survival time (in days) of heart attack patient

X_i = 1 if patient receives cardiac catheterization,

= 0 otherwise

- Clinical trials show that *CardCath* affects *SurvivalDays*.
- But is the treatment effective "in the field"?

Example: Cardiac Catheterization (2 of 3)

$$SurvivalDays_i = \beta_0 + \beta_1 CardCath_i + u_i$$

- Is OLS unbiased? The decision to treat a patient by cardiac catheterization is endogenous – it is (*was*) made in the field by EMT technician and depends on u_i (unobserved patient health characteristics)
- If healthier patients are catheterized, then OLS has simultaneous causality bias and OLS overstates overestimates the CC effect
- Propose instrument: distance to the nearest CC hospital minus distance to the nearest "regular" hospital

Example: Cardiac Catheterization (3 of 3)

- *Z* = differential distance to CC hospital
  - Relevant? If a CC hospital is far away, patient won' t bet taken there and won' t get CC
  - Exogenous? If distance to CC hospital doesn' t affect survival, other than through effect on $CardCath_i$, then corr(distance,u_i) = 0 so exogenous
  - If patients location is random, then differential distance is "as if" randomly assigned.
  - *The 1st stage is a linear probability model: distance affects the probability of receiving treatment*
- Results:
  - OLS estimates significant and large effect of CC
  - TSLS estimates a small, often insignificant effect

Example: Crowding Out of Private Charitable Spending (1 of 4)

Gruber, Jonathan and Daniel M. Hungerman (2005), "Faith-Based Charity and Crowd Out During the Great Depression," NBER Working Paper 11332.

Does government social service spending crowd out private (church, Red Cross, etc.) charitable spending?

Y = private charitable spending (churches)

X = government spending

What is the motivation for using instrumental variables?

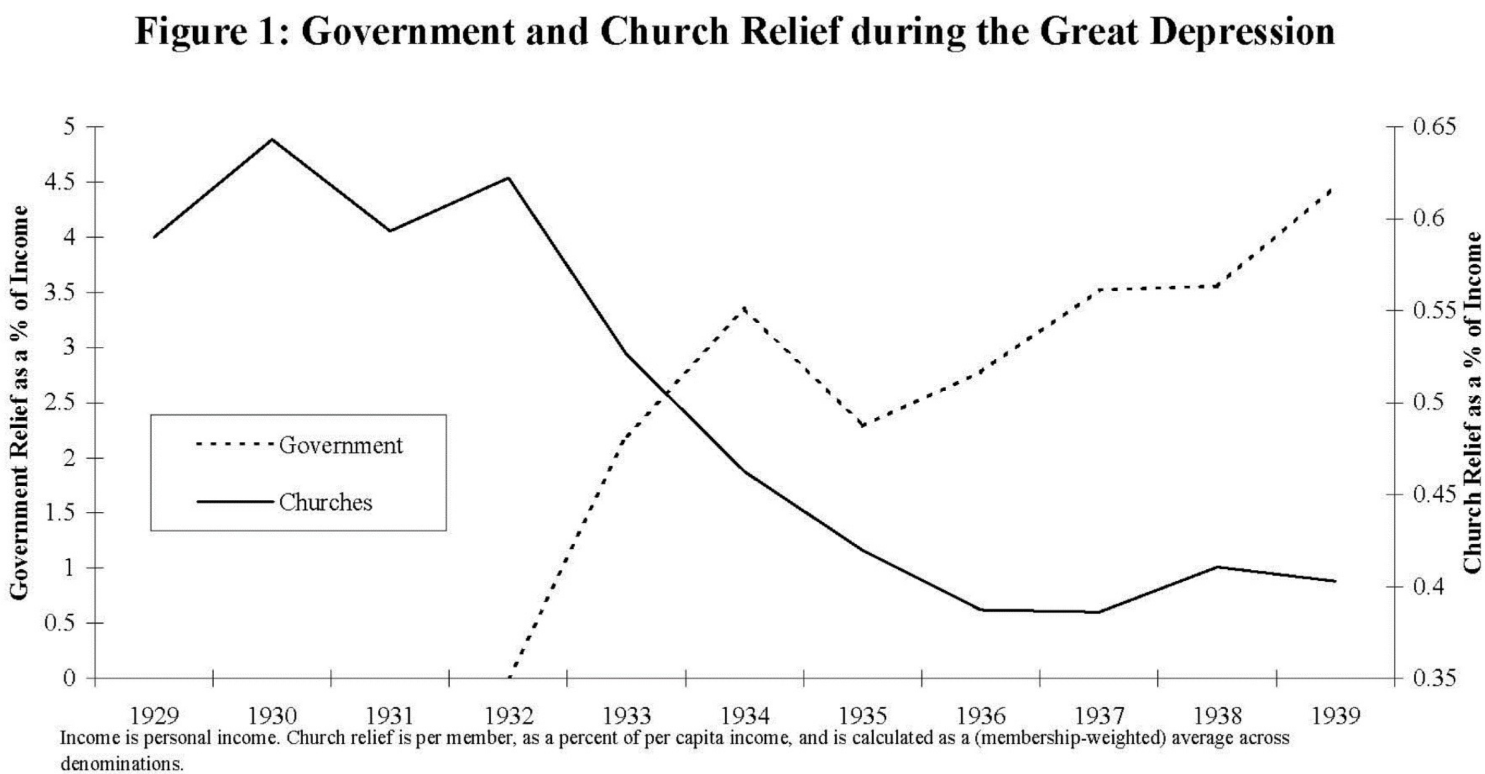
Proposed instrument:

Z = strength of Congressional delegation

Example: Crowding Out of Private Charitable Spending (2 of 4)

- panel data, yearly, by state, 1929–1939, U.S.
- *Y* = total benevolent spending by six church denominations (CCC, Lutheran, Northern Baptist, Presbyterian (2), Southern Baptist); benevolences = ¼ of total church expenditures.
- *X* = Federal relief spending under New Deal legislation (General Relief, Work Relief, Civil Works Administration, Aid to Dependent Children,⋯)
- *Z* = tenure of state’ s representatives on House & Senate Appropriations Committees, in months
- *W* = lots of fixed effects

Example: Crowding Out of Private Charitable Spending (3 of 4)



Example: Crowding Out of Private Charitable Spending (4 of 4)

Assessment of validity:

- Instrument validity:
  - Relevance?
  - Exogeneity?
- Other threats to internal validity:
  1. OV bias
  2. Functional form
  3. Measurement error
  4. Selection
  5. Simultaneous causality
- External validity to today in U.S.? To aid to developing countries?

Example: School Competition (1 of 3)

Hoxby, Caroline M. (2000), "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90, 1209–1238

What is the effect of public school competition on student performance?

Y = 12th grade test scores

X = measure of choice among school districts (function of

# of districts in metro area)

What is the motivation for using instrumental variables?

Proposed instrument:

Z = # small streams in metro area

Example: School Competition (2 of 3)

Data – some details

- cross-section, US, metropolitan area, late 1990s (n = 316),
- Y = 12th grade reading score (other measures too)
- X = index taken from industrial organization literature measuring the amount of competition ("Gini index") – based on number of "firms" and their "market share"
- Z = measure of small streams – which formed natural geographic boundaries.
- W = lots of control variables

Example: School Competition (3 of 3)

Assessment of validity:

- Instrument validity:
  - Relevance?
  - Exogeneity?
- Other threats to internal validity:
  1. OV bias
  2. Functional form
  3. Measurement error
  4. Selection
  5. Simultaneous causality
- External validity to today in U.S.? To aid to developing countries?

Conclusion (SW Section 12.6)

- A valid instrument lets us isolate a part of X that is uncorrelated with u, and that part can be used to estimate the effect of a change in X on Y
- IV regression hinges on having valid instruments:
  1. *Relevance*: Check via first-stage F
  2. *Exogeneity*: Test *over*identifying restrictions via the J-statistic
- A valid instrument isolates variation in X that is "as if" randomly assigned.
- The critical requirement of at least m valid instruments cannot be tested – *you must use your head.*

Some IV FAQs (1 of 2)

1. **When might I want to use IV regression?**

Any time that X is correlated with u and you have a valid instrument. The primary reasons for correlation between X and u could be:

- Omitted variable(s) that lead to OV bias
  - Ex: ability bias in returns to education
- Measurement error
  - Ex: measurement error in years of education
- Selection bias
  - Patients select treatment
- Simultaneous causality bias
  - Ex: supply and demand for butter, cigarettes

Some IV FAQs (2 of 2)

2. **What are the threats to the internal validity of an IV regression?**

- The main threat to the internal validity of IV is the failure of the assumption of valid instruments. Given a set of control variables W, instruments are valid if they are relevant and exogenous.
  - Instrument relevance can be assessed by checking if instruments are weak or strong: Is the first-stage F-statistic > 10?
  - Instrument exogeneity can be checked using the J-statistic – as long as you have m exogenous instruments to start with! In general, instrument exogeneity must be assessed using expert knowledge of the application.