

## Instrumental Variables Regression (SW Ch. 10)

EC 471  
Spring 2004

10-1

### The IV Estimator with a Single Regressor and a Single Instrument (SW Section 10.1)

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

- Loosely, 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  $\beta_1$ .
- This is done using an *instrumental variable*,  $Z_i$ , which is uncorrelated with  $u_i$ .
- The instrumental variable detects movements in  $X_i$  that are uncorrelated with  $u_i$ , and use these two estimate  $\beta_1$ .

10-3

Three important threats to intern

- simultaneous (causality) bias ( $X$  and  $Y$  are endogenous).
- errors-in-variables bias ( $X$  is
- unobserved omitted variable is correlated with  $X$  but is un included in the regression;

Then, we say that  $X_i$  and  $u_i$  are c

Instrumental variables regression these three sources.

### Terminology: endogeneity and

An *endogenous* variable is one t  
An *exogenous* variable is one th

*Historical note:* “Endogenous  
“determined within the system  
is jointly determined with  $Y$ ,  
to simultaneous causality. H  
narrow and IV regression can  
bias and errors-in-variable bi  
simultaneous causality bias.

## Two conditions for a valid instrument

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ where } Z_i \text{ is the IV for } X_i.$$

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

1. **Instrument relevance:**  $\text{corr}(Z_i, X_i) \neq 0$   
 $Z_i$  and  $X_i$  should be highly correlated.
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  $\beta_1$ ?

10-5

(2) Replace  $X_i$  by  $\hat{X}_i$  in the regression of interest:

regress  $Y$  on  $\hat{X}_i$  using OLS:

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

- Because  $\hat{X}_i$  is uncorrelated with  $u_i$  in large samples, so the first least squares assumption holds
- Thus  $\beta_1$  can be estimated by OLS using regression (2)
- This argument relies on large samples (so  $\pi_0$  and  $\pi_1$  are well estimated using regression (1))
- This the resulting estimator is called the “Two Stage Least Squares” (TSLS) estimator,  $\hat{\beta}_1^{TSLS}$ .

10-7

**The IV Estimator, one  $X$  and one  $Z$**   
Explanation #1: Two Stage Least Squares  
As it sounds, TSLS has two stages:

(1) First isolates the part of  $X$  that is correlated with  $Z$  by regressing  $X$  on  $Z$  using OLS

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

- Because  $Z_i$  is uncorrelated with  $u_i$ , we don’t need to worry about  $v_i$  being uncorrelated with  $u_i$ . We don’t need to estimate  $\pi_0$  and  $\pi_1$ , so...

- Compute the predicted value of  $X_i$  from the first stage regression:  $\hat{\pi}_0 + \hat{\pi}_1 Z_i, i = 1, \dots, n.$

Suppose you have a valid instrument  $Z_i$ .

Stage 1:

Regress  $X_i$  on  $Z_i$ , obtain the predicted values  $\hat{X}_i$ .

Stage 2:

Regress  $Y_i$  on  $\hat{X}_i$ ; the coefficient on  $\hat{X}_i$  is the TSLS estimator,  $\hat{\beta}_1^{TSLS}$ .

Then  $\hat{\beta}_1^{TSLS}$  is a consistent estimator of  $\beta_1$ .

Explanation #2: (only) a little algebra

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

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

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

10-9

### Consistency of the TOLS estimator

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

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

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

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

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

10-11

### The IV Estimator, one X and one Y

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

The IV estimator replaces these with sample covariances:

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

$s_{YZ}$  and  $s_{XZ}$  are the sample covariances

This is the TOLS estimator – just

### Example #1: Supply and demand

IV regression was originally developed to estimate supply and demand elasticities for agricultural products like butter:

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

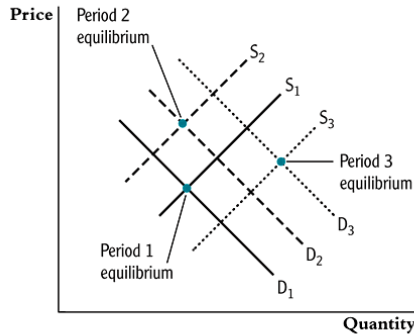
- $\beta_1$  = price elasticity of butter quantity for a 1% change in price (see specification discussion)
- Data: observations on price and quantity for different years
- The OLS regression of  $\ln(Q_i^{\text{butter}})$  from simultaneous causality

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

This interaction of demand and supply

**FIGURE 10.1**

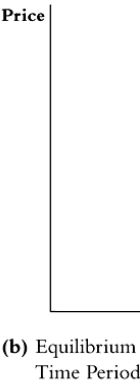
(a) Price and quantity are determined by the intersection of the supply and demand curves. The equilibrium in the first period is determined by the intersection of the demand curve  $D_1$  and the supply curve  $S_1$ . Equilibrium in the second period is the intersection of  $D_2$  and  $S_2$ , and equilibrium in the third period is the intersection of  $D_3$  and  $S_3$ .



(a) Demand and Supply in Three Time Periods

**FIGURE 10.1**

(b) This scatterplot shows equilibrium price and quantity in eleven different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



(b) Equilibrium Price and Quantity in Eleven Time Periods

Would a regression using these points estimate the demand curve?

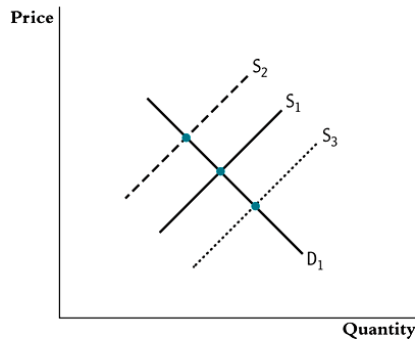
10-13

What would you get if only supply shifted?

TSLS in the supply-demand example

**FIGURE 10.1**

(c) When the supply curve shifts from  $S_1$  to  $S_2$  to  $S_3$  but the demand curve remains at  $D_1$ , the equilibrium prices and quantities trace out the demand curve.



(c) Equilibrium Price and Quantity When Only the Supply Curve Shifts

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

Let  $Z$  = rainfall in dairy-producing regions

Is  $Z$  a valid instrument?

(1) Exogenous?  $\text{corr}(rain_i, u_i) = 0$

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

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

*Plausibly*: insufficient rainfall means less butter

- 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.

10-15

## TSLS in the supply-demand example, ctd.

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

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

Stage 1: regress  $\ln(P_i^{butter})$  on  $rain$ , get  $\widehat{\ln(P_i^{butter})}$   
 $\widehat{\ln(P_i^{butter})}$  isolates changes in log price that arise from supply (part of supply, at least)

Stage 2: regress  $\ln(Q_i^{butter})$  on  $\widehat{\ln(P_i^{butter})}$   
The regression counterpart of using shifts in the supply curve to trace out the demand curve.

10-17

## Example: Demand for Cigarettes

- How much will a hypothetical cigarette tax reduce cigarette consumption?
- To answer this, we need the elasticity of demand for cigarettes, that is,  $\beta_1$ , in the regression,

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

- Will the OLS estimator plausibly be unbiased?

*Why or why not?*

10-19

## Inference using TSLS

$\hat{\beta}_1^{TSLS}$  is approx. distrib

- Statistical inference proceeds i
- This all assumes that the instru
- discuss what happens if they a
- **Important note on standard e**
  - The OLS standard errors fr
  - regression aren't right – the
  - the estimation in the first st
  - Instead, use a single special
  - computes the TSLS estimat
  - As usual, use heteroskedast

## Example: Cigarette demand, c

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

Panel data:

- Annual cigarette consumption (including tax)
- 48 continental US states, 198

Proposed instrumental variable:

- $Z_i =$  general sales tax per pac
- Is this a valid instrument?

(1) Relevant?  $\text{corr}(SalesTax_i,$

(2) Exogenous?  $\text{corr}(SalesTa$

For now, use data for 1995 only.

First stage OLS regression:

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

Second stage OLS regression:

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

Combined regression with correct, heteroskedasticity-robust standard errors:

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

(1.53) (0.32)

## Second stage

```
. reg lpackpc lravphat if year==1995, r;
```

Regression with robust standard errors

```
Number of obs = 48
F( 1, 46) = 10.54
Prob > F = 0.0022
R-squared = 0.1525
Root MSE = .22645
```

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lravphat	-1.083586	.3336949	-3.25	0.002	-1.755279	-.4118932
_cons	9.719875	1.597119	6.09	0.000	6.505042	12.93471

- These coefficients are the TSLS estimates
- The standard errors are wrong because they ignore the fact that the first stage was estimated

10-21

10-23

*STATA Example: Cigarette de*

Instrument =  $Z = rtaxso = gener$

```
. reg lragvprs rtaxso if year==1995, r;
```

Regression with robust standard errors

	Coef.	Robust Std. Err.	t
rtaxso	.0307289	.0048354	6.3
_cons	4.616546	.0289177	159.6

```
. predict lravphat;
```

Now we have the pre

Combined into a single command

```
. ivreg lpackpc (lravgprs = rtaxso) if year==1995, r;
```

IV (2SLS) regression with robust standard errors

	Coef.	Robust Std. Err.	t
lravgprs	-1.083587	.3189183	-3.4
_cons	9.719876	1.528322	6.3

Instrumented: lragvprs This is the  
Instruments: rtaxso This is the

OK, the change in the SEs was small th

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

(1.53) (0.32)

## Summary of IV Regression with a Single $X$ and $Z$

- A valid instrument  $Z$  must satisfy two conditions:
  - (1) *relevance*:  $\text{corr}(Z_i, X_i) \neq 0$
  - (2) *exogeneity*:  $\text{corr}(Z_i, u_i) = 0$
- 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

10-25

### Example: cigarette demand

- Another determinant of cigarette demand is income; omitting income could result in omitted variable bias
- Cigarette demand with one  $X$ , one  $W$ , and 2 instruments (2  $Z$ 's):

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

$Z_{1i}$  = general sales tax component only <sub>$i$</sub>

$Z_{2i}$  = cigarette-specific tax component only <sub>$i$</sub>

- Other  $W$ 's might be state effects and/or year effects (*in panel data, later...*)

10-27

## The General IV Reg

(SW Section

- So far we have considered IV regression with one endogenous regressor ( $X$ ) and one instrument ( $Z$ )
- We need to extend this to:
  - multiple endogenous regressors
  - multiple included exogenous regressors

These need to be included in the regression

  - multiple instrumental variables

More (relevant) instruments are needed to reduce the variance of TSLS: the  $R^2$  increases, so you have more power

### The general IV regression model

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

- $Y_i$  is the dependent variable
- $X_{1i}, \dots, X_{ki}$  are the endogenous regressors (correlated with  $u_i$ )
- $W_{1i}, \dots, W_{ri}$  are the **included exogenous regressors** (not correlated with  $u_i$ )
- $\beta_0, \beta_1, \dots, \beta_{k+r}$  are the unknown parameters
- $Z_{1i}, \dots, Z_{mi}$  are the  $m$  instrumental variables (**exogenous variables**)

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

We need to introduce some new concepts and to extend some old concepts to the general IV regression model:

- Terminology: *identification* and *overidentification*
- 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  $\beta_1, \dots, \beta_k$

10-29

### General IV regression: TSLS, 1 endogenous regressor

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

- Instruments:  $Z_{1i}, \dots, Z_{mi}$
- First stage
  - Regress  $X_1$  on *all* the exogenous regressors: regress  $X_1$  on  $W_1, \dots, W_r, Z_1, \dots, Z_m$  by OLS
  - Compute predicted values  $\hat{X}_{1i}, i = 1, \dots, n$
- Second stage
  - Regress  $Y$  on  $\hat{X}_1, W_1, \dots, W_r$  by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SEs* are wrong
- To get correct *SEs*, do this in a single step

10-31

The coefficients  $\beta_1, \dots, \beta_k$  are said

- **exactly identified** if  $m = k$ .  
There are just enough instruments  $Z_1, \dots, Z_m$  to identify  $\beta_1, \dots, \beta_k$ .
- **overidentified** if  $m > k$ .  
There are more than enough instruments  $Z_1, \dots, Z_m$ . If so, you can test whether the instruments are valid (a test of the "overidentification hypothesis" – we'll return to this later)
- **underidentified** if  $m < k$ .  
There are too few instruments  $Z_1, \dots, Z_m$ . If so, you need to get more instruments.

### Example: Demand for cigarettes

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

$Z_{1i}$  = general sales tax,  $i = 1, \dots, n$

$Z_{2i}$  = cigarette-specific tax,  $i = 1, \dots, n$

- Endogenous variable:  $\ln(P_i^{\text{cigarettes}})$
- Included exogenous variable:  $\ln(Q_i^{\text{cigarettes}})$
- Instruments (excluded endogenous variables):  $Z_1, Z_2$  (general sales tax, cigarette-specific tax)
- Is the demand elasticity  $\beta_1$  overidentified, or underidentified?

## Example: Cigarette demand, one instrument

```

      Y      W      X      Z
. ivreg lpackpc lperinc (lragvprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                       F( 2,      45) =      8.19
                                                       Prob > F      = 0.0009
                                                       R-squared     = 0.4189
                                                       Root MSE     = .18957
    
```

lpackpc	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
lragvprs	-1.143375	.3723025	-3.07	0.004	-1.893231 - .3935191
lperinc	.214515	.3117467	0.69	0.495	-.413375 .842405
_cons	9.430658	1.259392	7.49	0.000	6.894112 11.9672

```

Instrumented: lragvprs
Instruments:  lperinc rtaxso
              STATA lists ALL the exogenous regressors
              as instruments - slightly different
              terminology than we have been using
    
```

- Running IV as a single command yields correct *SEs*
- Use `, r` for heteroskedasticity-robust *SEs*

10-33

TSLs estimates,  $Z = \text{sales tax } (m = 1)$

$$\ln(\widehat{Q_i^{\text{cigarettes}}}) = 9.43 - 1.14 \ln(\widehat{P_i^{\text{cigarettes}}}) + 0.21 \ln(\text{Income}_i) \quad (1.26) \quad (0.37) \quad (0.31)$$

TSLs estimates,  $Z = \text{sales tax, cig-only tax } (m = 2)$

$$\ln(\widehat{Q_i^{\text{cigarettes}}}) = 9.89 - 1.28 \ln(\widehat{P_i^{\text{cigarettes}}}) + 0.28 \ln(\text{Income}_i) \quad (0.96) \quad (0.25) \quad (0.25)$$

- **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

10-35

## Example: Cigarette demand, two instruments

```

      Y      W      X      Z1
. ivreg lpackpc lperinc (lragvprs = rtaxso rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors
    
```

lpackpc	Coef.	Robust Std. Err.	t
lragvprs	-1.277424	.2496099	-5.12
lperinc	.2804045	.2538894	1.10
_cons	9.894955	.9592169	10.30

```

Instrumented: lragvprs
Instruments:  lperinc rtaxso rtaxso
              STATA lists ALL the exogenous regressors
              as instruments - slightly different
              terminology than we have been using
    
```

### General IV regression: TSLs with multiple endogenous regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \epsilon_i$$

- Instruments:  $Z_{1i}, \dots, Z_{mi}$
- Now there are  $k$  first stage regressions
  - Regress  $X_1$  on  $W_1, \dots, W_r, Z_1, \dots, Z_m$
  - Compute predicted values  $\hat{X}_1$
  - Regress  $X_2$  on  $W_1, \dots, W_r, Z_1, \dots, Z_m$
  - Compute predicted values  $\hat{X}_2$
  - Repeat for all  $X$ 's, obtaining  $\hat{X}_k$

## TSLS with multiple endogenous regressors, ctd.

- Second stage
  - Regress  $Y$  on  $\hat{X}_{1i}, \hat{X}_{2i}, \dots, \hat{X}_{ki}, W_1, \dots, W_r$  by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but  $SEs$  are wrong
- To get correct  $SEs$ , do this in a single step
- *What would happen in the second stage regression if the coefficients were underidentified (that is, if  $\#instruments < \#endogenous\ variables$ ); for example, if  $k = 2, m = 1$ ?*

10-37

## A “valid” set of instruments in the general case

The set of instruments must be relevant and exogenous:

1. Instrument relevance: *Special case of one  $X$*   
At least one instrument must enter the population counterpart of the first stage regression.
2. Instrument exogeneity  
*All* the instruments are uncorrelated with the error term:  $\text{corr}(Z_{1i}, u_i) = 0, \dots, \text{corr}(Z_{mi}, u_i) = 0$

10-39

## Sampling distribution of the general IV regression model

- Meaning of “valid” instrument
- The IV regression assumptions
- Implications: if the IV regression then the TSLS estimator is not inference (testing, confidence interval) usual

## “Valid” instruments in the general case

(1) General instrument relevance

- *General case, multiple  $X$ 's*  
Suppose the second stage regression using the predicted values from the first stage regression. Then: the multicollinearity in this (second stage) regression
- *Special case of one  $X$*   
At least one instrument must enter the population counterpart of the first stage regression

## Implications: Sampling distribution of TSLS

- If the IV regression assumptions hold, then the TSLS estimator is normally distributed in large samples.
- Inference (hypothesis testing, confidence intervals) proceeds as usual.
- Two notes about standard errors:
  - The second stage *SEs* are incorrect because they don't take into account estimation in the first stage; to get correct *SEs*, run TSLS in a single command
  - Use heteroskedasticity-robust *SEs*, for the usual reason.
- *All this hinges on having valid instruments...*

10-41

## Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor:

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

First stage regression:

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

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

10-43

## Checking Instrument (SW Section

Recall the two requirements for

1. *Relevance* (special case of or  
At least one instrument must  
counterpart of the first stage
2. *Exogeneity*  
**All** the instruments must be u  
error term:  $\text{corr}(Z_{1i}, u_i) = 0, \dots$

*What happens if one of these req  
satisfied? How can you check?*

## What are the consequences of

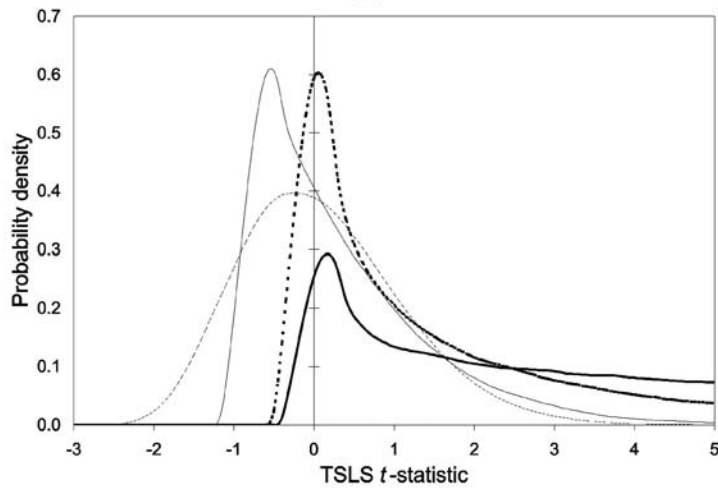
Consider the simplest case:

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

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

- The IV estimator is  $\hat{\beta}_1^{TSLS} = \frac{S_{YZ}}{S_{XZ}}$
- If  $\text{cov}(X, Z)$  is zero or small, the  
With weak instruments, the de
- If so, the sampling distribution  
statistic) is not well approxima  
approximation...

## An example: the distribution of the TSLS $t$ -statistic with weak instruments



Dark line = irrelevant instruments

Dashed light line = strong instruments

10-45

## Measuring the strength of instruments in practice: The first-stage $F$ -statistic

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

10-47

## Why does our trusty normal approximation fail?

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

- If  $\text{cov}(X, Z)$  is small, small changes in the sample to the next can induce large changes in the estimate.
- Suppose in one sample you calculate  $\hat{\beta}_1^{TSLS}$ .
- Thus the large- $n$  normal approximation to the sampling distribution is a poor approximation.
- A better approximation is that  $\hat{\beta}_1^{TSLS}$  is the *ratio* of two correlated normal variables (see SW App. 10.4)
- If instruments are weak, the usual normal approximation is unreliable – potentially very misleading.

## Checking for weak instruments

- 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 likely weak.***
- If so, the TSLS estimator will be biased and its inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
- Note that simply rejecting the null hypothesis that the coefficients on the  $Z$ 's are zero is not sufficient. You actually need substantial predictive power for the normal approximation to be a good approximation.
- There are more sophisticated tests for weak instruments. They compare  $F$  to 10 but they are based on different distributions.



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

### Distribution of the $J$ -statistic

- Under the null hypothesis that all the instruments are exogenous,  $J$  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.

10-53

### Panel data set

- Annual cigarette consumption, average prices paid by end consumer (including tax), personal income
- 48 continental US states, 1985-1995

### Estimation strategy

- Having panel data allows us to control for unobserved state-level characteristics that enter the demand for cigarettes, as long as they don't vary over time
- But we still need to use IV estimation methods to handle the simultaneous causality bias that arises from the interaction of supply and demand.

10-55

## Application to the Demand for Cigarettes (SW Section 10.4)

Why are we interested in knowing the demand for cigarettes?

- Theory of optimal taxation: cigarette demand elasticity: smaller deadweight loss, tax revenue affected less.
- Externalities of smoking – reasons for government intervention to discourage smoking
  - second-hand smoke (non-market)
  - monetary externalities

### Fixed-effects model of cigarette demand

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

- $i = 1, \dots, 48$ ,  $t = 1985, 1986, \dots$
- $\alpha_i$  reflects unobserved omitted variables for each state but not over time, e.g. state-level income
- Still,  $\text{corr}(\ln(P_{it}^{\text{cigarettes}}), u_{it})$  is probably non-zero due to the interaction of supply/demand interaction
- Estimation strategy:
  - Use panel data regression
  - Use TOLS to handle simultaneous causality bias

## Panel data IV regression: two approaches

- (a) The “ $n-1$  binary indicators” method
- (b) The “changes” method (when  $T=2$ )

### (a) The “ $n-1$ binary indicators” method

Rewrite

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

as

$$\ln(Q_{it}^{cigarettes}) = \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + \gamma_2 D2_{it} + \dots + \gamma_{48} D48_{it} + u_{it}$$

Instruments:

$$Z_{1it} = \text{general sales tax}_{it}$$

$$Z_{2it} = \text{cigarette-specific tax}_{it}$$

10-57

### (b) 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:

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

- Must 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
- We’ll take this approach

10-59

This now fits in the general IV r

$$\begin{aligned} \ln(Q_{it}^{cigarettes}) &= \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) \\ &\quad + \gamma_2 D2_{it} + \dots \end{aligned}$$

- $X$  (endogenous regressor) = 1
- 48  $W$ 's (included exogenous  $\ln(Income_{it}), D2_{it}, \dots, D48_{it}$ )
- Two instruments =  $Z_{1it}, Z_{2it}$
- Now estimate this full model
- An issue arises when dynamic adjustment is important, as in kick smoking – how to model

### STATA: Cigarette demand

#### First create “10-year change”

$$\begin{aligned} \text{10-year change in log price} &= \ln(P_{it}) - \ln(P_{it-10}) \end{aligned}$$

```
. gen dlpackpc = log(packpc/packpc[_n-10]);  
. gen dlavgprs = log(avgprs/avgprs[_n-10]);  
. gen dlperinc = log(perinc/perinc[_n-10]);  
. gen drtaxs = rtaxs-rtaxs[_n-10];  
. gen drtax = rtax-rtax[_n-10];  
. gen drtaxso = rtaxso-rtaxso[_n-10];
```

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

```
. ivreg Y W X Z
dlpackpc dlperinc (dlavgprs = drtaxso) , r;
```

IV (2SLS) regression with robust standard errors

Number of obs = 48  
 F( 2, 45) = 12.31  
 Prob > F = 0.0001  
 R-squared = 0.5499  
 Root MSE = .09092

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
dlpackpc						
dlavgprs	-.9380143	.2075022	-4.52	0.000	-1.355945	-.5200834
dlperinc	.5259693	.3394942	1.55	0.128	-.1578071	1.209746
_cons	.2085492	.1302294	1.60	0.116	-.0537463	.4708446

Instrumented: dlavgprs  
 Instruments: dlperinc drtaxso

**NOTE:**  
 - All the variables - Y, X, W, and Z's - are in 10-year changes  
 - Estimated elasticity = -.94 (SE = .21) - surprisingly elastic!  
 - Income elasticity small, not statistically different from zero  
 - Must check whether the instrument is relevant...

## What about two instruments (cig-only tax, sales tax)?

```
. ivreg dlpackpc dlperinc (dlavgprs = drtaxso drtax) , r;
```

IV (2SLS) regression with robust standard errors

Number of obs = 48  
 F( 2, 45) = 21.30  
 Prob > F = 0.0000  
 R-squared = 0.5466  
 Root MSE = .09125

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
dlpackpc						
dlavgprs	-1.202403	.1969433	-6.11	0.000	-1.599068	-.8057392
dlperinc	.4620299	.3093405	1.49	0.142	-.1610138	1.085074
_cons	.3665388	.1219126	3.01	0.004	.1209942	.6120834

Instrumented: dlavgprs  
 Instruments: dlperinc drtaxso drtax

drtaxso = general sales tax only  
 drtax = cigarette-specific tax only  
 Estimated elasticity is -1.2, even more elastic than using general sales tax only

## With $m > k$ , we can test the overidentifying restrictions

# Check instrument relevance:

```
. reg dlavgprs drtaxso dlperinc , r;
```

Regression with robust standard errors

	Coef.	Robust Std. Err.	t
dlavgprs			
drtaxso	.0254611	.0043876	5.8
dlperinc	-.2241037	.2188815	-1.0
_cons	.5321948	.0295315	18.0

```
. test drtaxso;
```

( 1) drtaxso = 0

F( 1, 45) = 33.67  
 Prob > F = 0.0000

First stage F = 33.7 > 10 so instr

## Can we check instrument exog

## Test the overidentifying restri

```
. predict e, resid; Computes predicted estimated regressio
```

```
. reg e drtaxso drtax dlperinc; Regress
```

Source	SS	df	MS
Model	.037769176	3	.012589725
Residual	.336952289	44	.007658007
Total	.374721465	47	.007972797

	Coef.	Std. Err.	t
e			
drtaxso	.0127669	.0061587	2.0
drtax	-.0038077	.0021179	-1.8
dlperinc	-.0934062	.2978459	-0.3
_cons	.002939	.0446131	0.0

```
. test drtaxso drtax;
```

( 1) drtaxso = 0

( 2) drtax = 0

F( 2, 44) = 2.47  
 Prob > F = 0.0966

Compute where F the inst

so J \*\* WARN

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

- $J = mF$ , where  $F$  = the  $F$ -statistic testing the coefficients on  $Z_{1i}, \dots, Z_{mi}$  in a regression of the TSLS residuals against  $Z_{1i}, \dots, Z_{mi}, W_{1i}, \dots, W_{mi}$ .
- Under the null hypothesis that all the instruments are exogenous,  $J$  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.
- In STATA:

```
. dis "J-stat = " r(df)*r(F) " p-value = " chiprob(r(df)-1,r(df)*r(F));
J-stat = 4.9319853 p-value = .02636401
```

$J = 2 \times 2.47 = 4.93$       p-value from chi-squared(1) distribution

10-65

Tabular summary of these results:

**TABLE 10.1** 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)
$\ln(D_{i,1995}^{cigarettes}) - \ln(D_{i,1985}^{cigarettes})$	-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	0.21 (0.13)	0.45** (0.14)	0.37** (0.12)
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 ten-year differences). The data are described in Appendix 10.1. The  $J$ -test of overidentifying restrictions is described in Key Concept 10.6 (its  $p$ -value is given in parentheses), and the first-stage  $F$ -statistic is described in Key Concept 10.5. Individual coefficients are statistically significant at the \*5% level or \*\*1% significance level.

10-67

## Check instrument relevance:

```
. reg dlvagprs X drtaxso Z1 drtax Z2 dlperinc W , r;
Regression with robust standard errors
```

	Coef.	Robust Std. Err.	t
dlvagprs			
drtaxso	.013457	.0031405	4.2
drtax	.0075734	.0008859	8.5
dlperinc	-.0289943	.1242309	-0.2
_cons	.4919733	.0183233	26.8

```
. test drtaxso drtax;
(1) drtaxso = 0
(2) drtax = 0
F( 2, 44) = 88.62      88.62
Prob > F = 0.0000
```

## How should we interpret the $J$ -test?

- $J$ -test rejects the null hypothesis that all instruments are exogenous
- This means that either  $rtaxso$  is endogenous, or both
- The  $J$ -test doesn't tell us which
- Why might  $rtax$  (cig-only tax) be endogenous?
  - Political forces: history of smoking ⇒ political pressure
  - If so, cig-only tax is endogenous
- This reasoning doesn't apply to  $rtaxso$
- ⇒ use just one instrument, the

## The Demand for Cigarettes: Summary of Empirical Results

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

10-69

### Remaining threats to internal validity, ctd.

- Remaining simultaneous causality bias?
  - Not if the general sales tax a valid instrument:
    - relevance?
    - exogeneity?
- Errors-in-variables bias? *Interesting question: are we accurately measuring the price actually paid? What about cross-border sales?*
- Selection bias? *(no, we have all the states)*

Overall, this is a credible estimate of the long-term elasticity of demand although some problems might remain.

10-71

## What are the remaining threats to internal validity?

- Omitted variable bias?
  - *Panel data estimator; proxy variables*
- Functional form mis-specification?
  - Hmm...should check...
    - A related question is the interpretation of the elasticity: using 10-year data, the interpretation is long-term elasticity. What would you obtain using shorter-run data?

## Where Do Valid Instruments Come From? (SW Section 10.4)

- Valid instruments are (1) relevant and (2) exogenous
- One general way to find instruments is to look for exogenous variation – variation that is randomly assigned in a randomized experiment
  - Rainfall shifts the supply curve, not the demand curve; rainfall is “as good as randomly assigned”
  - Sales tax shifts the supply curve, not the demand curve; sales tax is “as good as randomly assigned”
- Here is a final example...

## Example: Cardiac Catheterization

$$SurvivalDays_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

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”?

10-73

- $Z$  = differential distance to CC hospital
  - Relevant? If a CC hospital is far away, patient won't be taken there and won't get CC
  - Exogenous? If distance to CC hospital doesn't affect survival, other than through effect on *CardCath*, then  $\text{corr}(\text{distance}, u_i) = 0$  so exogenous
  - If patients location is random, then differential distance is “as if” randomly assigned.
  - *The 1<sup>st</sup> stage is a linear probability model: distance affects the probability of receiving treatment*
- Results (McClellan, McNeil, Newhous, *JAMA*, 1994):
  - OLS estimates significant and large effect of CC
  - TSLS estimates a small, often insignificant effect

10-75

- Is OLS unbiased? The decision to undergo cardiac catheterization is endogenous, made in the field by EMT technicians, not in a clinical trial (unobserved patient health characteristics).
- If healthier patients are catheterized, OLS overestimates the effect (simultaneous causality bias and omitted variables bias) – overestimates the CC effect
- Propose instrument: distance to nearest “regular” hospital – distance to the nearest “regular” hospital

## Summary: IV Regression (SW Section 10.4)

- A valid instrument lets us isolate the causal effect of a change in  $X$  on  $Y$ , uncorrelated with  $u$ , and that plausibly affects  $X$  to estimate the effect of a change in  $X$  on  $Y$ .
- IV regression hinges on having a valid instrument
  - (1) *Relevance*: check via first-stage regression
  - (2) *Exogeneity*: Test overidentified instruments via the  $J$ -statistic
- A valid instrument isolates variation in  $X$  that is randomly assigned.
- The critical requirement of at least one valid instrument cannot be tested – *you must use a valid instrument*