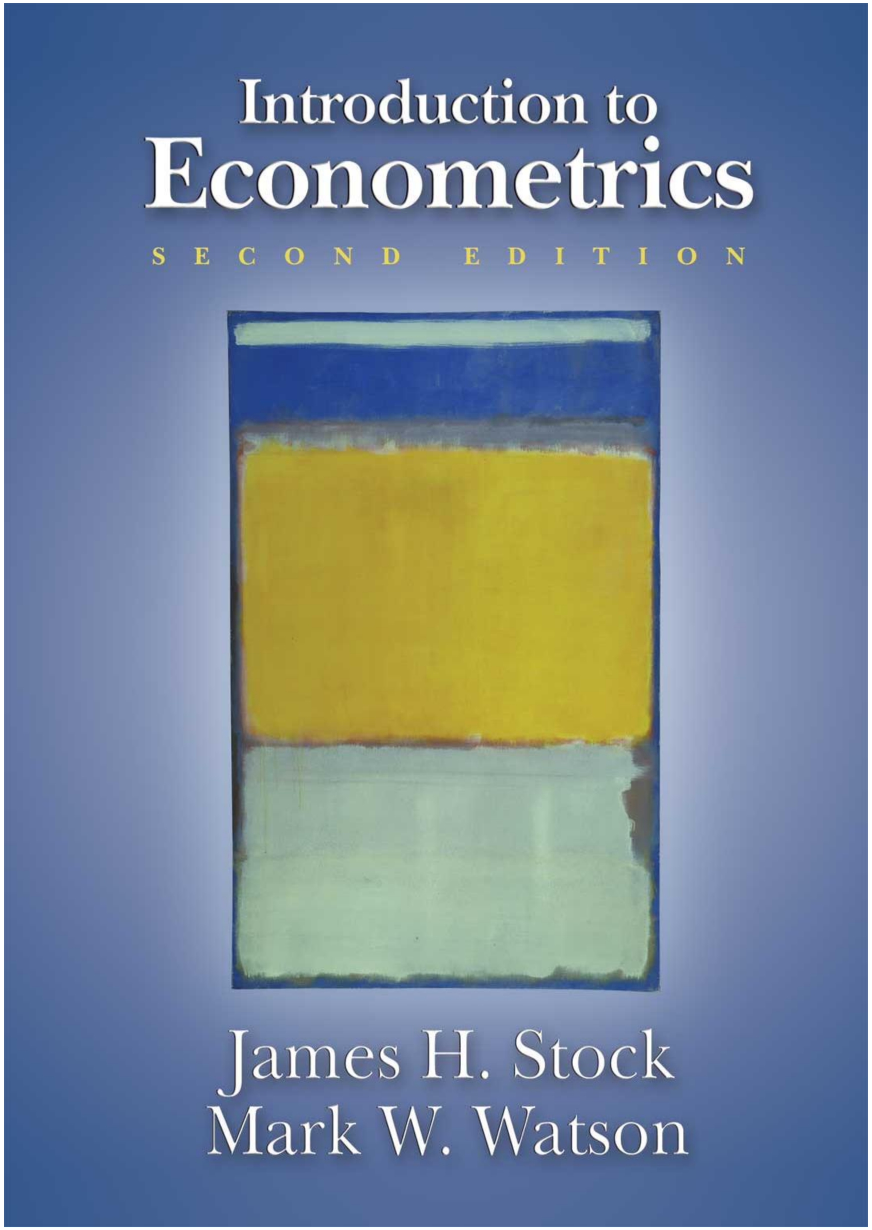
Chapter 12

Instrumental Variables Regression



Instrumental Variables Regression
(SW Chapter 12)

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;
- simultaneous causality bias (X causes Y, Y causes X);
- errors-in-variables bias (X is measured with error)

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

IV Regression with One Regressor and One Instrument (SW Section 12.1)

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

- 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 uncorrelated with u_i.
- The instrumental variable detects movements in X_i that are uncorrelated with u_i, and uses these to estimate β_1.

Terminology: endogeneity and exogeneity

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

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

Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality. However, this definition is narrow and IV regression can be used to address OV bias and errors-in-variable bias, not just to simultaneous causality bias.

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, one *X* and one *Z*

Explanation #1: Two Stage Least Squares (TSLS)

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

(1) First isolates the part of X that is uncorrelated with u: regress X on Z using OLS

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

- 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, $\hat{X}_i$, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

Two Stage Least Squares, ctd.

(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 (if n is large), the first least squares assumption holds (if n is large)**
- Thus β_1 can be estimated by OLS using regression (2)
- This argument relies on large samples (so π_0 and π_1 are well estimated using regression (1))
- This the resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares, ctd.

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, $\hat{\beta}_1^{TSLS}$.

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

The IV Estimator, one X and one Z, ctd.

Explanation #2: a little algebra…

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

Thus,

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

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

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

The IV Estimator, one X and one Z, ctd.

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

The IV estimator replaces these population covariances with sample covariances:

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

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

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

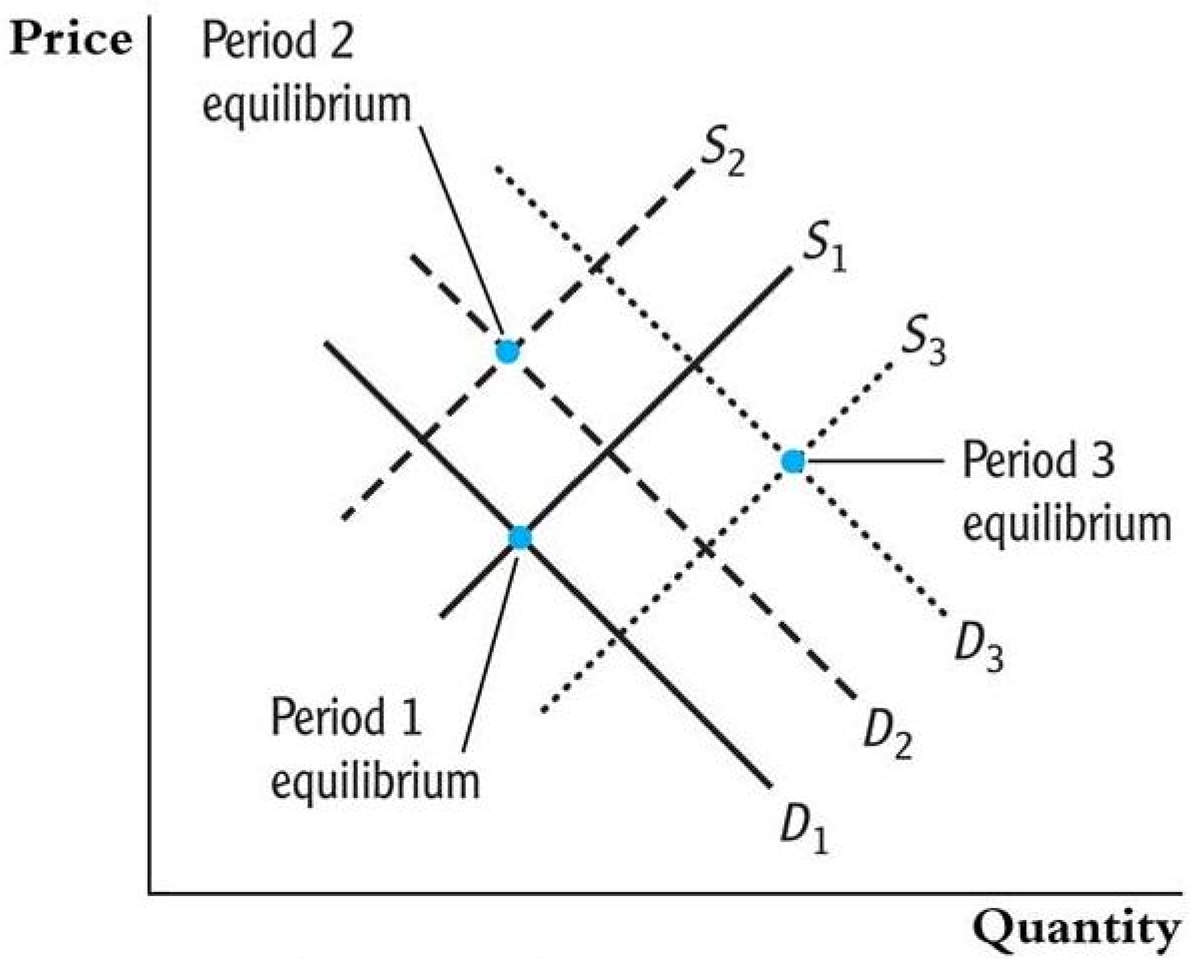
Example #1: Supply and demand for butter

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

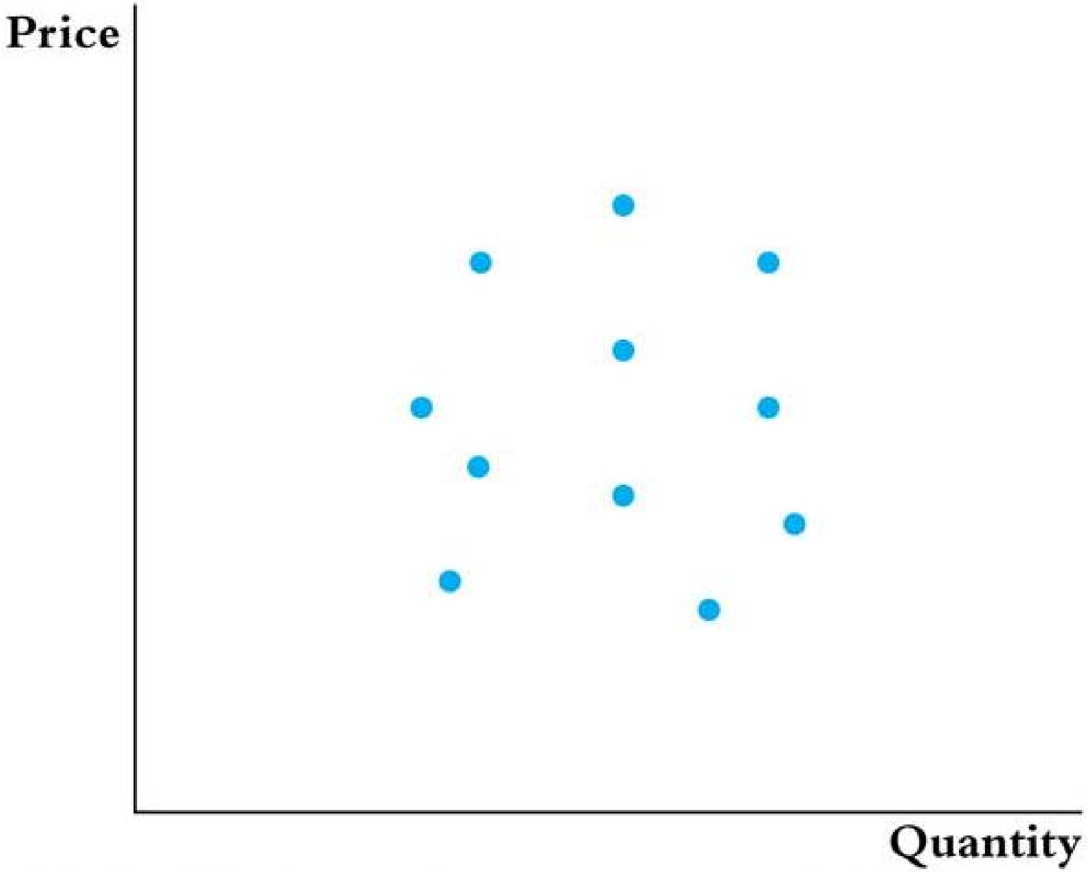
$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

- β_1 = price elasticity of 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*?)

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



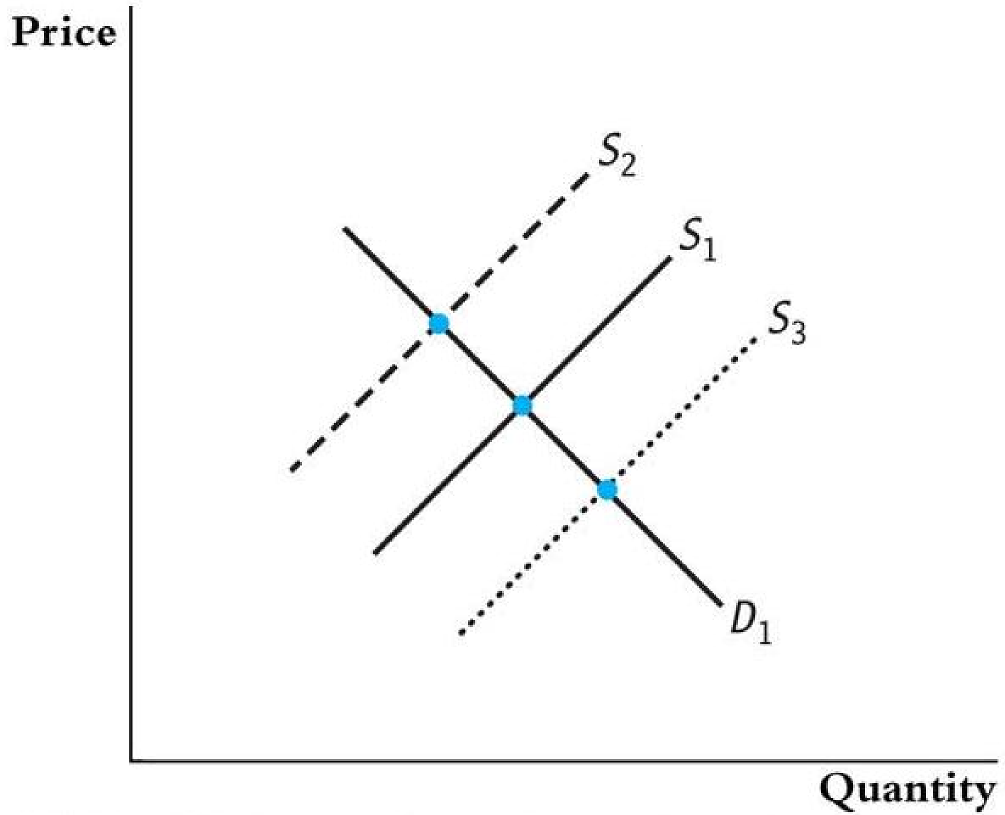
This interaction of demand and supply produces…



(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand cu

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:

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

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

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

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

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

Plausibly: insufficient rainfall means less grazing means less butter

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.

Example #2: Test scores and class size

- The California regressions still could have OV bias (e.g. parental involvement).
- This bias could be eliminated by using IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:

(1) relevant: $\text{corr}(Z_i, STR_i) \neq 0$

(2) exogenous: $\text{corr}(Z_i, u_i) = 0$

Example #2: Test scores and class size, ctd.

Here is a (hypothetical) instrument:

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

  $Z_i = Quake_i = 1$ if hit by quake, $= 0$ 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)

We'll go through other examples later…

Inference using TSLS

- In large samples, the sampling distribution of the TSLS estimator is normal
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. $\pm\ 1.96SE$
- 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 a sketch of the math (see SW App. 12.3 for the details)...

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

$$= \frac{\sum_{i=1}^{n} Y_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

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…

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

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 $\sqrt{n}$:

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

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

- $\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z) \neq 0$
- $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i$ is dist'd $N(0,\text{var}[(Z-\mu_Z)u])$ (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) ≠ 0 because the instrument is relevant

Inference using TSLS, ctd.

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

- 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: Cigarette demand, ctd.

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

Panel data:

- Annual cigarette consumption and average prices paid (including tax)
- 48 continental US states, 1985-1995

Proposed instrumental variable:

- Z_i = general sales tax per pack in the state = $SalesTax_i$
- Is this a valid instrument?

(1) Relevant? corr($SalesTax_i$, $\ln(P_i^{cigarettes})$) $\neq$ 0?

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

Cigarette demand, ctd.

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 regression with correct, heteroskedasticity-robust standard errors:

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

STATA Example: Cigarette demand, First stage

Instrument = Z = *rtaxso* = general sales tax (real $/pack)

```
             X         Z
. reg lravgprs rtaxso if year==1995, r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  1,    46) =   40.39
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4710
                                                       Root MSE      =  .09394

------------------------------------------------------------------------------
             |               Robust
    lravgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rtaxso |   .0307289   .0048354     6.35   0.000     .0209956    .0404621
       _cons |   4.616546   .0289177   159.64   0.000     4.558338    4.674755
------------------------------------------------------------------------------

              X-hat
. predict lravphat;       Now we have the predicted values from the 1st stage
```

Second stage

```
        Y       X-hat
. 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

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   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

Combined into a single command

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

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  1,    46) =   11.54
                                                       Prob > F      =  0.0014
                                                       R-squared     =  0.4011
                                                       Root MSE      =  .19035

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3189183    -3.40   0.001    -1.725536   -.4416373
       _cons |   9.719876   1.528322     6.36   0.000     6.643525    12.79623
------------------------------------------------------------------------------
Instrumented:  lravgprs          This is the endogenous regressor
Instruments:   rtaxso            This is the instrumental varible
------------------------------------------------------------------------------
```

OK, the change in the *SE*s was small *this time*...but not always!

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

$$(1.53) \quad (0.32)$$

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

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 ($X_1,\ldots,X_k$)
  - multiple included exogenous variables ($W_1,\ldots,W_r$)
    These need to be included for the usual OV reason
  - multiple instrumental variables ($Z_1,\ldots,Z_m$)
    More (relevant) instruments can produce a smaller variance of TSLS: the R^2 of the first stage increases, so you have more variation in $\hat{X}$.
- Terminology: identification & overidentification

Identification

- In general, a parameter is said to be ***identified*** if different values of the parameter would 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 $\beta_1,\ldots,\beta_k$
  - For example, suppose $k = 1$ but $m = 0$ (no instruments)!

Identification, ctd.

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

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

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

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

  There are more than enough instruments to estimate $\beta_1,\ldots,\beta_k$. *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 $\beta_1,\ldots,\beta_k$. *If so, you need to get more instruments!*

The general IV regression model: Summary of jargon

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

- Y_i is the ***dependent variable***
- $X_{1i},\ldots, X_{ki}$ are the ***endogenous regressors*** (potentially correlated with u_i)
- $W_{1i},\ldots,W_{ri}$ are the ***included exogenous variables*** or ***included exogenous regressors*** (uncorrelated with u_i)
- $\beta_0, \beta_1,\ldots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i},\ldots,Z_{mi}$ 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

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

- m instruments: $Z_{1i},\ldots, Z_m$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1,\ldots,W_r, Z_1,\ldots, Z_{\mathrm{m}}$ by OLS
  - Compute predicted values $\hat{X}_{1i}$, $i = 1,\ldots,n$
- Second stage
  - Regress Y on $\hat{X}_1$, $W_1,\ldots, W_r$ 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

Example: Demand for cigarettes

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

$$Z_{1i} = \text{general sales tax}_i$$
$$Z_{2i} = \text{cigarette-specific tax}_i$$

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

Example: Cigarette demand, one instrument

```
               Y       W         X          Z
. ivreg lpackpc lperinc (lravgprs = 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

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -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:  lravgprs
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 *SE*s
- Use , r for heteroskedasticity-robust *SE*s

Example: Cigarette demand, two instruments

```
                Y       W         X         Z1       Z2
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  2,    45) =   16.17
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4294
                                                       Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax   STATA lists ALL the exogenous regressors
                                     as "instruments" – slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

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

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

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

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

- 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

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

(1) ***Instrument exogeneity***: $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$

(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.

- Multicollinearity interpretation…
- *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

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

1. $E(u_i|W_{1i},\ldots,W_{ri}) = 0$
   - #1 says "the exogenous regressors are exogenous."
2. $(Y_i,X_{1i},\ldots,X_{ki},W_{1i},\ldots,W_{ri},Z_{1i},\ldots,Z_{mi})$ 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 $(Z_{1i},\ldots,Z_{mi})$ 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…

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:

   corr(Z_{1i},u_i) = 0,…, 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:

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

First stage regression:

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

- The instruments are relevant if at least one of $\pi_1,\ldots,\pi_m$ are nonzero.
- The instruments are said to be ***weak*** if all the $\pi_1,\ldots,\pi_m$ 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.

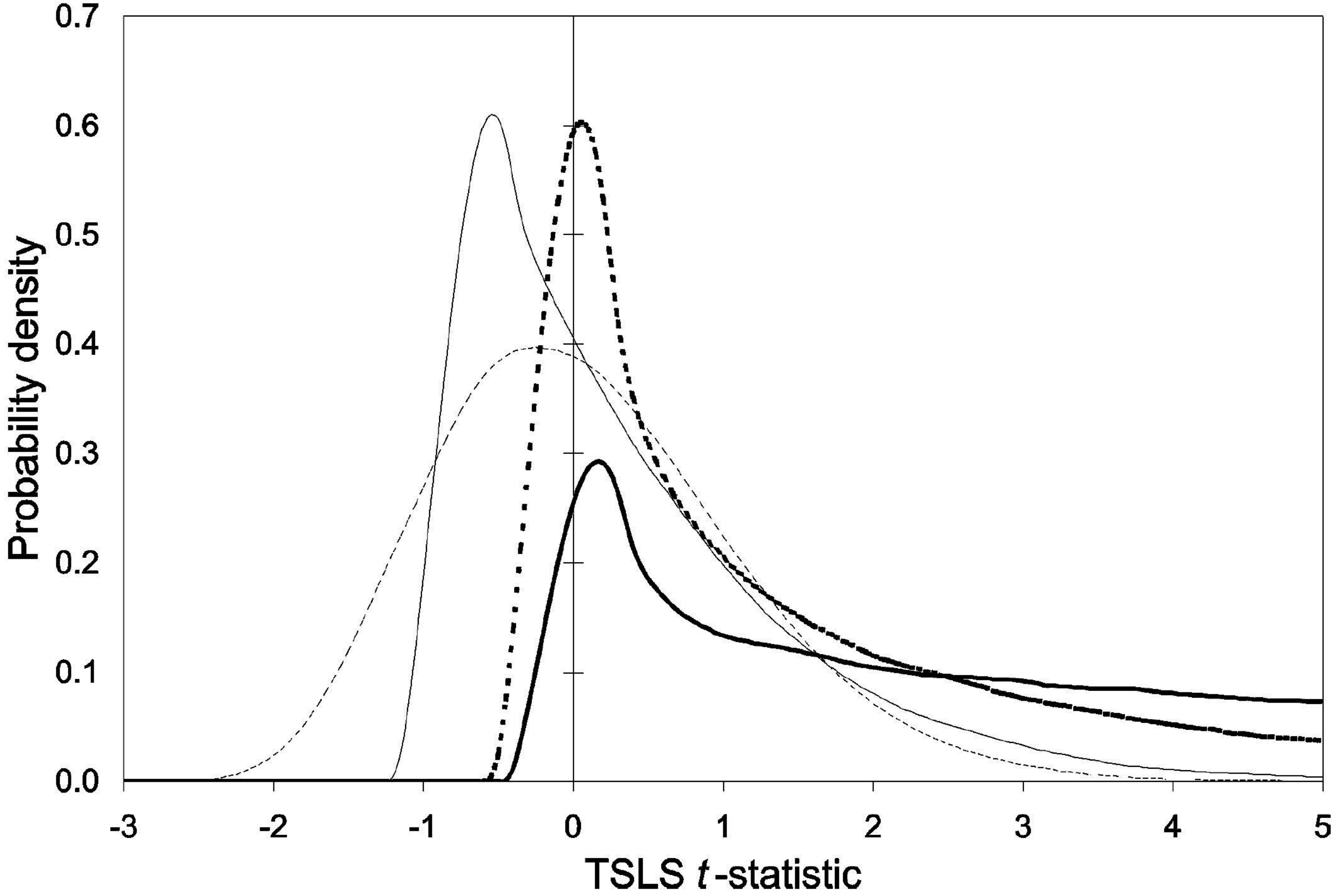
Consider the simplest case:

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

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

- 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 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 $Z_1,..,Z_m,W_1,\ldots,W_k$.

- Totally irrelevant instruments ⇔ *all* the coefficients on $Z_1,\ldots,Z_m$ are zero.
- The ***first-stage F-statistic*** tests the hypothesis that $Z_1,\ldots,Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

Checking for weak instruments with a single *X*

- 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.
- Note that simply rejecting the null hypothesis that the coefficients on the *Z*'s are zero isn't enough – you actually need substantial predictive content for the normal approximation to be a good one.
- There are more sophisticated things to do than just compare *F* to 10 but they are beyond this course.

What to do if you have weak instruments?

- Get better instruments (!)
- 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

- With weak instruments, TSLS conf. intervals are not valid – but some other confidence intervals *are*.
- The easiest of these is the Anderson-Rubin confidence interval, which is based on the Anderson-Rubin test statistic testing $\beta_1 = \beta_{1,0}$
  - Compute $Y_i^* = Y_i - \beta_{1,0}X_i$
  - Regress Y_i^* on $W_{1i},\ldots, W_{ri}, Z_{1i},\ldots, Z_{mi}$
  - The AR test is the F-statistic on $Z_{1i},\ldots, Z_{mi}$
- 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.
- This is valid even if instruments are irrelevant!
- Computation: should use specialized software…

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 does 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(!)
  - 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 doesn't successfully 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).

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

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

The *J*-test of overidentifying restrictions

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 $\hat{u}_i$ against $Z_{1i},\ldots,Z_{mi}, W_{1i},\ldots,W_{ri}$
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i},\ldots,Z_{mi}$ are all zero;
5. The ***J-statistic*** is $J = mF$

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

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

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.
   - 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

2. *Exogeneity*

- ***All*** the instruments must be uncorrelated with the error term: corr(Z_{1i},u_i) = 0,…, corr(Z_{mi},u_i) = 0
- We can partially test for exogeneity: if m>1, we can test the hypothesis that all are exogenous, against the alternative that as many as m–1 are endogenous (correlated with u)
- The test is the J-test, constructed using the TSLS residuals.

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: optimal tax is inverse to elasticity: smaller deadweight loss if quantity is affected less.
- Externalities of smoking – role for government intervention to discourage smoking
  - second-hand smoke (non-monetary)
  - monetary externalities

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.

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,…,48, t = 1985, 1986,…,1995
- α_i reflects unobserved omitted factors that vary across states but not over time, e.g. attitude towards smoking
- Still, corr($\ln(P_{it}^{cigarettes}), u_{it}$) is plausibly nonzero because of supply/demand interactions
- 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:

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

- 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

STATA: Cigarette demand

First create "10-year change" variables

10-year change in log price

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

```
. gen dlpackpc = log(packpc/packpc[_n-10]);    _n-10 is the 10-yr lagged value
. 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

```
               Y            W        X            Z
. ivreg 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

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    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…*

Check instrument relevance: compute first-stage *F*

```
.  reg dlavgprs drtaxso dlperinc , r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  2,    45) =   16.84
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5146
                                                       Root MSE      =  .06334

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0254611   .0043876     5.80   0.000      .016624    .0342982
    dlperinc |  -.2241037   .2188815    -1.02   0.311    -.6649536    .2167463
       _cons |   .5321948   .0295315    18.02   0.000     .4727153    .5916742
------------------------------------------------------------------------------

.  test drtaxso;

 ( 1)  drtaxso = 0

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

We didn't need to run "test" here because with m=1 instrument, the F-statistic is the square of the t-statistic, that is, 5.80*5.80 = 33.67

First stage F = 33.7 > 10 so instrument is not weak

Can we check instrument exogeneity? **No**: $m = k$

Check instrument relevance: compute first-stage *F*

```
                X         Z1         Z2       W
.  reg dlavgprs drtaxso drtax dlperinc , r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  3,    44) =   66.68
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7779
                                                       Root MSE      =  .04333

------------------------------------------------------------------------------
             |               Robust
    dlavgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |    .013457   .0031405     4.28   0.000     .0071277    .0197863
       drtax |   .0075734   .0008859     8.55   0.000     .0057879    .0093588
    dlperinc |  -.0289943   .1242309    -0.23   0.817    -.2793654    .2213767
       _cons |   .4919733   .0183233    26.85   0.000     .4550451    .5289015
------------------------------------------------------------------------------

.  test drtaxso drtax;

 ( 1)  drtaxso = 0
 ( 2)  drtax = 0

       F(  2,    44) =   88.62             88.62 > 10 so instruments aren't weak
            Prob > F =    0.0000
```

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

------------------------------------------------------------------------------
             |               Robust
    dlpackpc |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    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…

Test the overidentifying restrictions

```
.  predict e, resid;      Computes predicted values for most recently
                              estimated regression (the previous TSLS regression)
.  reg e drtaxso drtax dlperinc;      Regress e on Z's and W's

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =    1.64
       Model |  .037769176     3  .012589725           Prob > F      =  0.1929
    Residual |  .336952289    44  .007658007           R-squared     =  0.1008
-------------+------------------------------           Adj R-squared =  0.0395
       Total |  .374721465    47  .007972797           Root MSE      =  .08751

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0127669   .0061587     2.07   0.044      .000355    .0251789
       drtax |  -.0038077   .0021179    -1.80   0.079     -.008076    .0004607
    dlperinc |  -.0934062   .2978459    -0.31   0.755    -.6936752    .5068627
       _cons |    .002939   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------
.  test drtaxso drtax;

 ( 1)  drtaxso = 0                        Compute J-statistic, which is m*F,
 ( 2)  drtax = 0                          where F tests whether coefficients on
                                              the instruments are zero

       F(  2,    44) =    2.47                so J = 2 × 2.47 = 4.93
            Prob > F =    0.0966          ** WARNING – this uses the wrong d.f. **
```

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

- $J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\ldots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\ldots,Z_{mi}, W_{1i},\ldots,W_{mi}$.
- 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.
- 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 × 2.47 = 4.93 *p*-value from chi-squared(1) distribution

Now what???

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)	−1.34** (0.23)	−1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage *F*-statistic	33.70	107.20	88.60
Overidentifying restrictions *J*-test and *p*-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the ten-year differences). The data are described in Appendix 12.1. The *J*-test of overidentifying restrictions is described in Key Concept 12.6 (its *p*-value is given in parentheses), and the first-stage *F*-statistic is described in Key Concept 12.5. Individual coefficients are statistically significant at the *5% level or **1% significance level.

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 $\Rightarrow$ political pressure for low cigarette taxes
  - If so, cig-only tax is endogenous
- This reasoning doesn't apply to general sales tax
- $\Rightarrow$ 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:

  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?
   - *Panel data estimator; probably OK*
2. Functional form mis-specification (*could check this*)
3. Remaining simultaneous causality bias?
   - Not if the general sales tax a valid instrument:
     - relevance? exogeneity?
4. Errors-in-variables bias?
5. Selection bias? (*no, we have all the states*)

External validity?

- *This is a long-run elasticity*

Finding IVs: Examples (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

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"?

Cardiac catheterization, ctd.

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

- Is OLS unbiased? The decision to treat a patient by cardiac catheterization is endogenous – it is (*was*) made in the field by EMT technician 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

Cardiac catheterization, ctd.

- 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

Gruber, Jonathan and Danniel 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

Private charitable spending, ctd.

Data – some details

- 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); benevolances = 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

Private charitable spending, ctd.

Figure 1: Government and Church Relief during the Great Depression

Income is personal income. Church relief is per member, as a percent of per capita income, and is calculated as a (membership-weighted) average across denominations.

Private charitable spending, ctd.

***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

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

School competition, ctd.

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

School competition, ctd.

***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?

Summary: IV Regression

(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. 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

2. What is the list of threats to the internal validity of an IV regression?

- IV regression is internally valid under the IV regression assumptions.

- The threats to the internal validity of IV are cases in which the IV regression assumptions do not hold. The key assumptions are:
  - $E(u_i|W_{1i},\ldots, W_{ri}) = 0$ (W's are exogenous)
  - IVs are valid (relevant and exogenous)

Threats to internal validity of IV, ctd.

- The W's might not be exogenous for the usual 5 reasons. In many IV regressions, the exogeneity of the W's is plausible.
- The Z's might not be valid instruments. The two main threats to the validity of IV regression thus are:
  1. Weak instruments
  2. The instruments are not exogenous (corr(Z_i,u_i) $\neq$ 0).

3. Why is this list different from the list of 5 threats for multiple regression?

Because the main threat to the internal validity of IV regression is having invalid instruments.