Chapter 12:
Instrumental Variables Regression
Outline

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

For the simple model \( Y_i = \beta_0 + \beta_1 X_i + u_i \) the OLS estimate of \( \beta_1 \) is

\[
\hat{\beta}_1 = \beta_1 + \frac{S_{X,u}}{S_X^2}
\]

(see p.141, eqn. 4.30)

So it is “off from the truth” \( \beta_1 \) by this error.

Now, in ch4 we saw how OLS assn’s 1&2 imply \( E[\text{ratio}] = 0 \) so OLS is unbiased in that \( E(\hat{\beta}_1) = \beta_1 \).

But it is common (see ch9’s threats) for these OLS assn’s to be false and for OLS to be biased.

The problem that concerns us here is the implication that OLS is not consistent (since unbiased): As samples increase, the error does not converge to 0.

The error converges to \( \frac{\text{cov}(X,u)}{\text{var}(X)} \)
Gist of IV: Correct X so OLS is consistent

- IV solution “corrects” X to $\tilde{X}$ & u to $\tilde{u}$ so that:
  1. the coefficients stay the same as in model: $Y_i = \beta_0 + \beta_1 \tilde{X} + \tilde{u}$
  2. the corresponding error does converge to zero: $\frac{s_{\tilde{X},\tilde{u}}}{s_{\tilde{X}}^2} \to 0$

That is, after the correction:
  1. **Parameters of interest** stay the same in new model
  2. OLS estimate of new model is consistent

**Gist:** One should regress Y on the corrected X instead!

Note: (2) holds $\iff$ error’s limit $\frac{\text{cov}(\tilde{X}, \tilde{u})}{\text{var}(\tilde{X})}$ is 0: $\text{cov}(\tilde{X}, \tilde{u}) = 0$
Rationale for jargon of “instrument”

• In practice, we use auxiliary variable(s) \( Z \) to “build the correction” of \( X \) to \( \tilde{X} \)

• When we use a drill/hammer in a building project, we call these tools/instruments. So here, \( Z \) is ...

• known as an **instrument** or **instrumental variable** or **IV**.

• For **any** \( \text{rv} \) \( Z \), one can decompose \( X \) as \( X = \pi_0 + \pi_1 Z + \nu \) and call \( \tilde{X} := \pi_0 + \pi_1 Z \) a “correction.”

• Whether \( Z \) “works” (makes \( \tilde{X} \) satisfy 1&2) is the matter!
IV Estimator w/ Single Regressor & Instrument

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

- For any rv \( Z \), decompose \( X \) as \( X = \pi_0 + \pi_1 Z + w \)
  subject to the restrictions \( E(w) = 0, \text{cov}(Z, w) = 0 \) (two equations)
  This is possible: 2 equations in 2 unknowns \( \pi \).

Exercise: Show solution is \( \tilde{\pi}_1 = \frac{\text{cov}(X, Z)}{\text{var}(Z)}, \tilde{\pi}_0 = E(X) - \pi_1 E(Z) \)

- Define \( \tilde{X} := \tilde{\pi}_0 + \tilde{\pi}_1 Z \), so \( X = \tilde{X} + w \)
  Recall, we seek \( Z \) so that both conditions hold:

1. \( Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i \)
2. \( \text{cov}(\tilde{X}, \tilde{u}) = 0 \)
What $Z$ to choose? What makes a good IV?

**Claim:** For any rv $Z$, condition 1 holds with $\tilde{u}_i := \beta_1 w + u_i$

Proof: From original model,

$$Y_i = \beta_0 + \beta_1 X + u_i = \beta_0 + \beta_1 (\tilde{X} + w) + u_i = \beta_0 + \beta_1 \tilde{X} + (\beta_1 w + u_i)$$

Note, we used only $X = \tilde{X} + w$, nothing about $\pi$.

**Claim:** Sps. $\text{cov}(Z,X) \neq 0$. Then condition 2 holds with given $\pi$ if and only if $\text{cov}(Z,u) = 0$.

Proof: The 2$^{nd}$ condition is

$$0 = \text{cov}(\tilde{X}, \tilde{u}) = \text{cov}(\tilde{\pi}_0 + \tilde{\pi}_1 Z, \beta_1 w + u) = \pi_1 \text{cov}(Z, \beta_1 w + u) = \tilde{\pi}_1 [\beta_1 0 + \text{cov}(Z,u)] = \tilde{\pi}_1 \text{cov}(Z,u)$$

Now, $\tilde{\pi}_1 := \frac{\text{cov}(X,Z)}{\text{var}(Z)}$ by assumption is not zero, so equation holds if and only if $\text{cov}(Z,u) = 0$
Two key conditions

• A rv $Z$ “works” (makes $\tilde{X}$ meet both conditions) if and only if $\text{cov}(Z,u)=0$ (“exogeneity”), assuming $\text{cov}(X,Z)\neq 0$

• We certainly want $\text{cov}(X,Z)\neq 0$ to hold (“relevance”) because otherwise $\tilde{X} = \tilde{\pi}_0 = E(Z)$ and $Y_i = (\beta_0 + \beta_1 E(Z)) + \tilde{u}_i$

That is, there is only an intercept, and though it can be estimated, the contribution of each $\beta$ cannot be sorted out!

We summarize these conditions in the following definition:

*Def.* A random variable $Z$ is a valid instrument for $Y_i = \beta_0 + \beta_1 X_i + u_i$ if it is exogenous: $\text{cov}(Z,u)=0$ relevant: $\text{cov}(Z,X)\neq 0$
Conclusion & TSLS

If \( Z \) is a valid instrument for the model \( Y_i = \beta_0 + \beta_1 X_i + u_i \), then the OLS estimates of the regression \( Y_i = \beta_0 + \beta_1 \tilde{X}_i + v_i \) converge to the original parameters of interest.

**Theoretical IV:** One should regress \( Y \) on the correction \( X \).

Wrinkle: We don’t know what the corrected \( X \) is!

\[ \tilde{X}_i := \tilde{\pi}_0 + \tilde{\pi}_1 Z_i \quad \text{where} \quad \tilde{\pi}_1 := \frac{\text{cov}(Z, X)}{\text{var}(Z)} \quad \text{and} \quad \tilde{\pi}_0 := E(X) - \tilde{\pi}_1 E(Z) \]

Wrinkle: We don’t know \( \tilde{\pi} \) and its population E’s and cov’s. Solution: Use known sample analogues, thus defining **TSLS**:

1. OLS-regress \( X \) on instrument to get *predicted correction*
2. OLS-regress \( Y \) on *predicted correction* (That is, replace unknown \( \tilde{\pi} \) by known OLS \( \hat{\pi} \)).
Single instrument & regressor: Formula for slope coefficient

There is an expression for the slope coefficient in terms of the instrument’s “validity,” ie cov(Z,Y) and cov(Z,X):

\[
\text{cov}(Z,Y) = \text{cov}(Z, \beta_0 + \beta_1 X + u) = \beta_1 \text{cov}(Z, X) + \text{cov}(Z,u)
\]

By instrument relevance, \( \text{cov}(Z,X) \neq 0 \), and by exogeneity, \( \text{cov}(Z,u) = 0 \), so upon division

\[
\beta_1 = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,X)}
\]

Compare this to the classical formula \( \frac{\text{cov}(X,Y)}{\text{cov}(X,X)} \) – in this, \( Z=X \) is valid because of the OLS ass’n 1.

The new formula suggests that if the instrument is weak (\( \text{cov}(Z,X) \) is close to zero), then estimates of \( \beta_1 \) will be unreliable (to be seen)
Mechanically, TSLS is clear from its definition:
1. Regress $X$ on instrument to get predicted correction $\hat{X}$
2. Regress $Y$ on predicted correction $\hat{X}$

Now we derive a formula for the resulting estimate.
From (2), the estimate of $Y = \gamma_0 + \gamma_1 \hat{X} + \nu_i$ is $\hat{\gamma}_1 := \frac{s_{Y,\hat{X}}}{s_{\hat{X}}^2}$
From (1), predicted correction is $\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z$ with $\hat{\pi}_1 = \text{OLS} \frac{s_{X,Z}}{s_Z^2}$
Substituting and using covariance-rules,
$$\hat{\gamma}_1 = \frac{\hat{\pi}_1 s_{Y,Z}}{\hat{\pi}_1^2 s_Z^2} = \frac{1}{\hat{\pi}_1} \frac{s_{Y,Z}}{s_X^2} \frac{s_Z^2}{s_{X,Z}^2} = \frac{s_{Y,Z}}{s_{X,Z}}$$

Compare to previous slide:
TSLS’s estimate is sample analogue of theoretical IV’s estimate. Since sample moments converge to population moments, and theoretical IV is consistent, **TSLS is consistent!**
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.
- If irrelevant, then dividing by \( s_{XZ} \) very random.
Terminology: Endogeneity & Exogeneity

An **endogenous** variable is one that is correlated with $u$
An **exogenous** variable is one that is uncorrelated with $u$

IV regression addresses an endogenous $X$ with an exogenous instrumental variable $Z$.

*Digression on terminology:* “Endogenous” literally means “determined within the system.” If $X$ is jointly determined with $Y$, then a regression of $Y$ on $X$ is subject to simultaneous causality bias. But this definition of endogeneity is too narrow because IV regression can be used to address OV bias and errors-in-variable bias. Thus we use the broader definition of endogeneity above.
Example #1: Effect of Studying on Grades

What is the effect on grades of studying one extra hour per day?

\[ Y = \text{GPA} \]
\[ X = \text{study time (hours per day)} \]

Data: grades and study hours of college freshmen.

Would you expect the OLS estimator of \( \beta_1 \) to be unbiased? Why or why not?

What is students of high ability (positive \( u \)) enjoy more studying (high \( X \))? Then \( X \) is endogenous; OLS#1 false.
Studying on grades, ctd.


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

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

1. Is it relevant (correlated with $X$)?
2. Is it exogenous (uncorrelated with $u$)?
Studying on grades, ctd.

\[ X = \pi_0 + \pi_1 Z + \nu_i \]
\[ Y = \gamma_0 + \gamma_1 Z + w_i \ldots \text{reduced form} \]
\[ Y = \text{GPA (4 point scale)} \]
\[ X = \text{time spent studying (hours per day)} \]
\[ Z = 1 \text{ if roommate brought video game, } = 0 \text{ otherwise} \]

Stinebrinckner and Stinebrinckner’s findings

\[ \hat{\pi}_1 = -0.668 \]
\[ \hat{\gamma}_1 = -0.241 \]
\[ \hat{\beta}_{IV}^1 = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-0.241}{-0.668} = 0.360 \]

What are the units? Do these estimates make sense in a real-world way? (Note: They actually ran the regressions including additional regressors – more on this later.)
Example #2: Supply & demand for butter

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

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

- $\beta_1$ = price elasticity of butter = percent change in quantity for a 1% change in price (recall log-log specification discussion)

- Data: observations on price & quantity of butter for different years

- The OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ suffers from simultaneous causality bias ... price causes quantity, quantity causes price
Simultaneous causality bias in the OLS regression of $\ln(Q_{i}^{\text{butter}})$ on $\ln(P_{i}^{\text{butter}})$ arises because price and quantity are determined by the interaction of demand and supply:

(a) Demand and supply in three time periods
Interaction of demand and supply produces data like…

(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?
But...what if only supply were to shift?

- 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 \( Z = \text{rainfall in dairy-producing regions} \).

Is \( Z \) a valid instrument?

1. **Relevant?** \( \text{corr}(\text{rain}_i,\ln(P_i^{butter})) \neq 0? \)
   
   *Plausibly*: insufficient rainfall means less grazing means less butter means higher prices

2. **Exogenous?** \( \text{corr}(\text{rain}_i,u_i) = 0? \)
   
   *Plausibly*: whether it rains in dairy-producing regions shouldn’t affect demand for butter
TSLS in the supply-demand example, ctd.

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

\[ Z_i = \text{rain}_i = \text{rainfall in dairy-producing regions.} \]

Stage 1: regress \( \ln(P_i^{butter}) \) on \( \text{rain} \), get \( \ln(P_i^{butter}) \) isolates changes in log price affecting only supply

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

The regression counterpart of using shifts in the supply curve to trace out the demand curve.
Example #3:  Test scores and class size

• The California test score/class size regressions still could have OV bias (e.g. parental involvement).

• In principle, this bias can be eliminated by IV regression (TSLS).

• IV regression requires a valid instrument, that is, an instrument that is:
  1. relevant: \( \text{corr}(Z_i, STR_i) \neq 0 \)
  2. exogenous: \( \text{corr}(Z_i, u_i) = 0 \)
Example #3: Test scores and class size, ctd.

Here is a (hypothetical) instrument:
- some districts, randomly hit by an earthquake, “double up” classrooms:
  \[ Z_i = Quake_i = 1 \text{ if hit by quake, } = 0 \text{ otherwise} \]
- *Do the two conditions for a valid instrument hold?*
- The earthquake makes it *as if* the districts were in a random assignment experiment. Thus, the variation in *STR* arising from the earthquake is exogenous.
- The first stage of TSLS regresses *STR* against *Quake*, thereby isolating the part of *STR* that is exogenous (the part that is “as if” randomly assigned)
Inference using TSLS

• 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 (Appendix 12.3 has details)
\[
\hat{\beta}_{1}^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Z_i - \bar{Z}) \div \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z})
\]

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

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

\[ \hat{\beta}_{1}^{TSLS} - \beta_1 = \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})} \approx \frac{1}{n} \sum_{i=1}^{n} u_i (Z_i - \mu_Z) \frac{1}{n} \sum_{i=1}^{n} X_i (Z_i - \bar{Z}). \]

Now, the denominator converges in probability to \( \text{cov}(X, Z) \).

The numerator is an average of iid, zero expectation rvs \((X, Z \text{ uncorrelated})\), each with variance \( \text{var}(u(Z-\mu)) \). By the CLT, the average has \( 1/n \) of this variance. So

\[ \beta_1^{TSLS} \approx N(0, \nu) \text{ with } \nu = \frac{1}{n} \frac{\text{var}(u_i(Z_i - \mu_Z))}{\left(\text{cov}(Z_i, X_i)\right)^2}. \]
Inference using TSLs, ctd.

\[ \hat{\beta}_{1}^{TLS} \text{ is approx. distributed } N(\beta_1, \sigma^2_{\beta_{TLS}}), \]

- 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 shortly what happens if they aren’t.

**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 SEs.
- As usual, use heteroskedasticity-robust SEs.
Example #4: Demand for Cigarettes

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

Why is the OLS estimator of \( \beta_1 \) likely to be biased?

• Data set: Panel data on annual cigarette consumption and average prices paid (including tax), by state, for the 48 continental US states, 1985-1995.

• Proposed instrumental variable:
  • \( Z_i = \) general sales tax per pack in the state = \( SalesTax_i \)
  • Do you think this instrument is plausibly valid?
    1. Relevant? \( \text{corr}(SalesTax_i, \ln(P_i^{cigarettes})) \neq 0? \)
    2. Exogenous? \( \text{corr}(SalesTax_i, u_i) = 0? \)
Cigarette demand, ctd.

For now, use data from 1995 only. First stage OLS regression:
\[
\ln(P_{i}^{\text{cigarettes}}) = 4.63 + .031 SalesTax_{i}, \ n = 48
\]

Second stage OLS regression:
\[
\ln(Q_{i}^{\text{cigarettes}}) = 9.72 - 1.08 \ln(P_{i}^{\text{cigarettes}}), \ n = 48
\]

Combined TSLS regression with correct, heteroskedasticity-robust standard errors:
\[
\ln(Q_{i}^{\text{cigarettes}}) = 9.72 - 1.08 \ln(P_{i}^{\text{cigarettes}}), \ n = 48
\]
\[
(1.53) \ (0.32)
\]
**STATA Example: Cigarette demand, First stage**

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

\[
\begin{align*}
X & Z \\
\text{. reg lravgprs rtaxso if year==1995, r;}
\end{align*}
\]

Regression with robust standard errors

\[
\begin{array}{c}
\text{Number of obs} = 48 \\
F(1, 46) = 40.39 \\
\text{Prob > F} = 0.0000 \\
\text{R-squared} = 0.4710 \\
\text{Root MSE} = 0.09394
\end{array}
\]

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>lravgprs</td>
<td>Coef. Std. Err. t P&gt;</td>
</tr>
<tr>
<td>rtaxso</td>
<td>0.0307289 0.0048354 6.35 0.000 0.0209956 0.0404621</td>
</tr>
<tr>
<td>_cons</td>
<td>4.616546 0.0289177 159.64 0.000 4.558338 4.674755</td>
</tr>
</tbody>
</table>

\[
X-hat \\
\text{. predict lravphat; Now we have the predicted values from the 1st stage}
\]
Second stage

\[ Y \quad X-hat \]

. 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 = 0.22645

|             | Robust
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
</tr>
<tr>
<td>lpackpc</td>
<td></td>
</tr>
<tr>
<td>lravphat</td>
<td>-1.083586</td>
</tr>
<tr>
<td>_cons</td>
<td>9.719875</td>
</tr>
</tbody>
</table>

- 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:

\[
Y \quad X \quad Z
\]

. `ivregress 2sls lpackpc (lravgprs = rtaxso) if year==1995, vce(robust);`

Instrumental variables (2SLS) regression

|                 | Coef.   | Std. Err. |      z  |   P>|z|   | [95% Conf. Interval] |
|-----------------|---------|-----------|---------|---------|---------------------|
| lpackpc         | -1.083587 | .3122035 | -3.47   | 0.001   | -1.695494, -0.471679 |
| _cons           | 9.719876  | 1.496143  | 6.50    | 0.000   | 6.78749, 12.65226   |

Instrumented: lravgprs  This is the endogenous regressor
Instruments: rtaxso     This is the instrumental variable

Estimated cigarette demand equation:

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

\[
(1.53) \quad (0.31)
\]
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

- 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 \( (W_1, \ldots, W_r) \) or control variables, which 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 \( \hat{X} \).

- *New terminology*: identification & overidentification
Identification

- In general, a parameter is said to be **identified** if different values of the parameter produce different distributions of the data.

- In IV regression, whether the coefficients are identified depends on the relation between the number of instruments ($m$) and the number of endogenous regressors ($k$).

- Intuitively, if there are fewer instruments than endogenous regressors, we can’t estimate $\beta_1, \ldots, \beta_k$.
  - 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”) ...

- **underidentified** if $m < k$.
  
  There are too few instruments to estimate $\beta_1, \ldots, \beta_k$. If so, you need 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 regressors** (uncorrelated with \( u_i \)) or **control variables** (included so that \( Z_i \) is uncorrelated with \( u_i \), once the \( W \)'s are included)
- \( \beta_0, \beta_1, \ldots, \beta_{k+r} \) 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

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 w_{1i} + \cdots + \beta_{1+r} w_{ri} + u_i \]

- \( m \) instruments: \( z_{1i}, \ldots, z_m \)
- First stage
  - Regress \( x_1 \) on all the exogenous regressors: regress \( x_1 \) on \( w_{1r}, \ldots, w_r, z_{1r}, \ldots, z_m \), and an intercept, by OLS
  - Compute predicted values \( \hat{x}_{1i} \), \( i = 1, \ldots, n \)
- Second stage
  - Regress \( y \) on \( \hat{x}_{1i}, w_{1r}, \ldots, w_r \), and an intercept, by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but \( SEs \) are wrong
- To get correct \( SEs \), take single step in Stata
Example #4: Demand for cigarettes, ctd.
Suppose income is exogenous (is this plausible?), and we also want to estimate the income elasticity:

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

We actually have two instruments:

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

- Endogenous variable: \(\ln(P_i^{\text{cigarettes}})\) (“one \(X\)’’)
- Included exogenous variable: \(\ln(\text{Income}_i)\) (“one \(W\)’’)
- Instruments (excluded endogenous variables): general sales tax, cigarette-specific tax (“two \(Zs\)’’)
- \(Is \beta_1 \text{ over–, under–, or exactly identified? ‘‘Over’’!}\)
**Example: Cigarette demand, one instrument**

IV: rtaxso = real overall sales tax in state

\[ Y = W \times X + Z \]

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

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

|                       | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----------------------|--------|-----------|-------|------|----------------------|
| lpackpc               |        |           |       |      |                      |
| lravgprs             | -1.143375 | .3723025 | -3.07 | 0.004 | -1.893231 to -.3935191 |
| lperinc              | .214515  | .3117467 | 0.69  | 0.495 | -0.413375 to .842405  |
| _cons                | 9.430658 | 1.259392 | 7.49  | 0.000 | 6.894112 to 11.9672   |

**Number of obs = 48**

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F( 2, 45)</td>
<td>8.19</td>
<td>Prob &gt; F</td>
<td>0.0009</td>
<td>R-squared</td>
<td>0.4189</td>
<td>Root MSE</td>
</tr>
</tbody>
</table>

**Root MSE**

\[ .18957 \]

**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 the correct SEs
- Use , r for heteroskedasticity-robust SEs
**Example:** Cigarette demand, **two** instruments

\[ Y \quad W \quad X \quad Z_1 \quad Z_2 \]

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

IV (2SLS) regression with robust standard errors

Number of obs = 48

| Coef.    | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------|-----------|------|------|---------------------|
| lpackpc  | -1.277424 | .2496099 | -5.12 | 0.000 | -1.780164, -.7746837 |
| lravgprs | .2804045  | .2538894  | 1.10  | 0.275 | -.230955, .7917641   |
| lperinc  | 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, $Z = \text{sales tax (} m = 1 \text{)}$

\[
\ln(Q_{i}^{\text{cigarettes}}) = 9.43 - 1.14 \ln(P_{i}^{\text{cigarettes}}) + 0.21 \ln(\text{Income}_i)
\]

(1.26) (0.37) (0.31)

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

\[
\ln(Q_{i}^{\text{cigarettes}}) = 9.89 - 1.28 \ln(P_{i}^{\text{cigarettes}}) + 0.28 \ln(\text{Income}_i)
\]

(0.96) (0.25) (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
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*

Let \( \hat{X}_j^* \) be the prediction from a *population* regression of \( X_j \) (the jth regressor) on all instruments \( Z \) & included exogenous regressors \( W \). The requirement is that these not be collinear:

\( (\hat{X}_1^*, \ldots, \hat{X}_k^*, W_1, \ldots, W_r, 1) \)

*Note: This prediction is infeasible in practice (“population”)*

*Special case of one X*: the general assumption is equivalent to (a) at least one instrument must have nonzero coefficient in 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. \( E(u_i|W_{1i},\ldots,W_{ri}) = 0 \) ... allows \( X,u \) to be correlated
   • #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
**W’s as control variables**

• In many cases, the purpose of including the W’s is to control for omitted factors, so that once the W’s are included, Z is uncorrelated with u. If so, W’s don’t need to be exogenous; instead, the W’s need to be effective control variables in the sense discussed in Chapter 7 – except now with a focus on producing an exogenous instrument.

• Technically, the condition for W’s being effective control variables is that the conditional mean of $u_i$ does not depend on $Z_i$, given $W_i$:

$$E(u_i|W_i, Z_i) = E(u_i|W_i)$$
W’s as control variables, ctd.

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

\[ E(u_i|W_i, Z_i) = E(u_i|W_i) \]

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

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

• For details, see SW Appendix 12.6
Example #1: Effect of studying on grades, ctd

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

\( Y \) = first-semester GPA
\( X \) = average study hours per day
\( Z = 1 \) if roommate brought video game, \( = 0 \) otherwise
Roommates were randomly assigned

Can you think of a reason that \( Z \) might be correlated with \( u \) – even though it is randomly assigned? What else enters the error term – what are other determinants of grades, beyond time spent studying?
Example #1: Effect of studying on grades, ctd

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Why might \( Z \) be correlated with \( u \)?

- Here’s a *hypothetical* possibility: gender. Suppose:
  - Women get better grades than men, holding constant hour spent studying
  - Men are more likely to bring a video game than women
  - Then corr(\( Z_i, u_i \)) < 0 (males are more likely to have a [male] roommate who brings a video game – but males also tend to have lower grades, holding constant the amount of studying).

- This is just a version of OV bias. The solution to OV bias is to control for (or include) the OV – in this case, gender.
Example #1: Effect of studying on grades, ctd

• This logic leads you to include $W = \text{gender}$ as a control variable in the IV regression:

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

• The TSLS estimate reported above is from a regression that included gender as a $W$ variable – along with other variables such as individual $i$’s major.
Checking Instrument Validity

Recall the two requirements for valid instruments:

1. **Relevance** (special case of one X)
   At least one instrument must have a nonzero coefficient in a regression of X on instruments and exogenous regressors

2. **Exogeneity**
   All the instruments must be uncorrelated with the error term: \( \text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0 \)

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

*If you have multiple instruments, which should you use?*
Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor:

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

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 \( \hat{\beta}_{TSLS}^{T} = \frac{S_{YZ}}{S_{XZ}} \)
- If \( \text{cov}(X,Z) \) is zero or small, then \( s_{XZ} \) will be small: With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of \( \hat{\beta}_{TSLS}^{T} \) (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?

\[ \hat{\beta}_{1}^{\text{TSLS}} = \frac{S_{YZ}}{S_{XZ}} \]

- If \( \text{cov}(X,Z) \) is small, small changes in \( s_{XZ} \) (from one sample to the next) can induce big changes in \( \hat{\beta}_{1}^{\text{TSLS}} \).
- Suppose in one sample you calculate \( s_{XZ} = 0.0001... \).
- Thus the large-\( n \) normal approximation is a poor approximation to the sampling distribution of \( \hat{\beta}_{1}^{\text{TSLS}} \).
- A better approximation is that \( \hat{\beta}_{1}^{\text{TSLS}} \) 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 $Z_1, \ldots, Z_m, W_1, \ldots, W_k$.
- Totally irrelevant instruments $\iff$ 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.

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

- If so, the TSLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
Checking for Weak Instruments with a Single X, ctd.

- Why compare the first-stage $F$ to 10?

- Simply rejecting the null hypothesis that the coefficients on the $Z$’s are zero isn’t enough – you need substantial predictive content for the normal approximation to be good.

- Comparing the first-stage $F$ to 10 tests for whether the bias of TSLS, relative to OLS, is less than 10%. If $F$ is smaller than 10, the relative bias exceeds 10%—that is, TSLS can have substantial bias (see SW App. 12.5).
What to do if you have weak instruments

• Get better instruments (if possible!)
• 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 $\beta_1$ and construction of confidence intervals
  – This seems odd, but TSLS isn’t normally distributed...
Skip: Confidence Intervals with Weak Instruments (Appendix 12.5)

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

1. The Anderson-Rubin confidence interval

   - The Anderson-Rubin confidence interval is based on the Anderson-Rubin test statistic testing $\beta_1 = \beta_{1,0}$:
     - Compute $c_i = Y_i - \beta_{1,0}X_i$
     - Regress 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 $\beta_1$ not rejected at the 5% level by the AR test.

   - Computation: should use specialized software.
2. Moreira’s Conditional Likelihood Ratio confidence interval

- The Conditional Likelihood Ratio (CLR) confidence interval is based on inverting Moreira’s Conditional Likelihood Ratio test. Computing this test, its critical value, and the CLR confidence interval requires specialized software.

- The CLR confidence interval tends to be tighter than the Anderson-Rubin confidence interval, especially when there are many instruments.

- If your software produces the CLR confidence interval, this is the one to use.
Skip: Estimation with Weak Instruments

There are no unbiased estimators if instruments are weak or irrelevant. However, some estimators have a distribution more centered around $\beta_1$ than TSLS.

- One such estimator is the limited information maximum likelihood estimator (LIML)
- The LIML estimator
  - can be derived as a maximum likelihood estimator
  - is the value of $\beta_1$ that minimizes the $p$-value of the AR test(!)

- For more discussion about estimators, tests, and confidence intervals when you have weak instruments, see SW, App. 12.5
Checking Assumption #2: Instrument Exogeneity

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

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

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

Consider the simplest case:

\[ Y_i = \beta_0 + \beta_1 X_i + u_{ir} \]

• Suppose there are two valid instruments: \( Z_{1i}, Z_{2i} \)
• Then you could compute two separate TSLS estimates.
• Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
• The \( J \)-test of overidentifying restrictions makes this comparison in a statistically precise way.
• This can only be done if \( \#Z's > \#X's \) (overidentified).
The J-test of Overidentifying Restrictions

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

\[ Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{ri} + u_i \]

The J-test:

1. First run above regression, focus on the residuals:
2. Compute the residuals \( \hat{u}_i = Y_i - \hat{Y}_i \)
3. Regress residuals against \( Z_{1i},...,Z_{mi}, W_{1i},...,W_{ri} \)
4. Compute the \( F \)-statistic testing the hypothesis that the coefficients on \( Z_{1i},...,Z_{mi} \) are all zero in the latter;
5. The **J-statistic** is \( J = mF \)

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

\[ J = mF, \] where $F = \text{the } F\text{-statistic testing the coefficients on } Z_{1i}, \ldots, Z_{mi} \text{ 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

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

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

• All the instruments must be uncorrelated with the error term: \( \text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0 \)

• We can partially test for exogeneity: if \( m > 1 \), we can test the null hypothesis that all the instruments are exogenous, against the alternative that as many as \( m-1 \) are endogenous (correlated with \( u \))

• The test is the \( J \)-test, which is constructed using the TSLS residuals.

• If the \( J \)-test rejects, then at least some of your instruments are endogenous – so you must make a difficult decision and jettison some (or all) of your instruments.
Why are we interested in knowing the elasticity of demand for cigarettes?

• Theory of optimal taxation. The optimal tax rate is inversely related to the price elasticity: the greater the elasticity, the less quantity is affected by a given percentage tax, so the smaller is the change in consumption and deadweight loss.

• Externalities of smoking – role for government intervention to discourage smoking
  – health effects of second-hand smoke? (non-monetary)
  – monetary externalities
Panel data set

- Annual cigarette consumption, average prices paid by end consumer (including tax), personal income, and tax rates (cigarette-specific and general statewide sales tax rates)

Estimation strategy

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

\[
\ln(Q_{it}^{cigarettes}) = \alpha_i + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(\text{Income}_{it}) + u_{it}
\]

- \( i = 1, \ldots, 48 \), \( t = 1985, 1986, \ldots, 1995 \)
- \( \text{corr}(\ln(P_{it}^{cigarettes}), u_{it}) \) is plausibly nonzero because of supply/demand interactions
- \( \alpha_i \) reflects unobserved omitted factors that vary across states but not over time, e.g. attitude towards smoking
- Estimation strategy:
  - Use panel data regression methods to eliminate \( \alpha_i \)
  - Use TSLS to handle simultaneous causality bias
The “changes” method (when $T=2$)

• One way to model long-term effects is to consider 10-year changes, between 1985 and 1995
• Rewrite the regression in “changes” form:
  \[
  \ln(Q_{i1995}) - \ln(Q_{i1985}) = \beta_1\left[\ln(P_{i1995}) - \ln(P_{i1985})\right] + \beta_2[\ln(Income_{i1995}) - \ln(Income_{i1985})] + (u_{i1995} - u_{i1985})
  \]

• Create “10-year change” variables, for example:
• 10-year change in log price = $\ln(P_{i1995}) - \ln(P_{i1985})$
• Then estimate the demand elasticity by TSLS using 10-year changes in the instrumental variables
• This is equivalent to using the original data and including the state binary indicators (“$W$” variables) in the regression
**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]); 
. gen dlavgprs = log(avgprs/avgprs[_n-10]); 
. gen delperinc = 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

\[
\begin{array}{c|ccc}
Y & W & X & Z \\
\end{array}
\]

. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso), r;

IV (2SLS) regression with robust standard errors

<table>
<thead>
<tr>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>dlpackpc</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>dlavgprs</td>
</tr>
<tr>
<td>dlperinc</td>
</tr>
<tr>
<td>_cons</td>
</tr>
</tbody>
</table>

Instrumented: dlavgprs
Instruments: dlperinc drtaxso

NOTE:
- All the variables - Y, X, W, and Z’s - are in 10-year changes
- Estimated elasticity = -0.94 (SE = 0.21) - 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;

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>.191437213</td>
<td>2</td>
<td>.095718606</td>
<td>F( 2, 45) = 23.86</td>
</tr>
<tr>
<td>Residual</td>
<td>.180549989</td>
<td>45</td>
<td>.004012222</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>.371987202</td>
<td>47</td>
<td>.007914621</td>
<td>R-squared = 0.5146</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.4931</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = .06334</td>
</tr>
</tbody>
</table>

| dlavgprs | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|----------|-------|-----------|------|-----|----------------------|
| drtaxso  | .0254611 | .0037374 | 6.81 | 0.000 | .0179337 - .0329885 |
| dlperinc | -.2241037 | .2119405 | -1.06 | 0.296 | -.6509738 - .2027664 |
| _cons    | .5321948 | .031249 | 17.03 | 0.000 | .4692561 - .5951334 |

. test drtaxso;
( 1) drtaxso = 0

We didn’t need to run “test” here!
With m=1 instrument, the $F$-stat is the square of the $t$-stat:

$$F( 1, 45) = 46.41$$
$$Prob > F = 0.0000$$
$$6.81 \times 6.81 = 46.41$$

First stage $F = 46.5 > 10$ so instrument is not weak

Can we check instrument exogeneity? **No**: $m = k$
Cigarette demand, 10 year changes – 2 IVs

\[ Y = W + X + Z_1 + Z_2 \]

```
.ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso drtax) , vce(r);
```

Instrumental variables (2SLS) regression

|               | Coef.   | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|---------------|---------|-----------|-------|-------|----------------------|
| dlpackpc      |         |           |       |       |                      |
| dlavgprs      | -1.202403 | .1906896  | -6.31 | 0.000 | -1.576148            | -.8286588           |
| dlperinc      | .4620299 | .2995177  | 1.54  | 0.123 | -.1250139            | 1.049074            |
| _cons         | .3665388 | .1180414  | 3.11  | 0.002 | .1351819             | .5978957            |

Instrumented: dlavgprs
Instruments: dlperinc drtaxso drtax

\[ \text{drtaxso} = \text{general sales tax only} \]
\[ \text{drtax} = \text{cigarette-specific tax only} \]

Estimated elasticity is -1.2, even more elastic than using general sales tax only!
First-stage $F$ – both instruments

\[
X \quad Z1 \quad Z2 \quad W
\]

. reg dlavgprs drtaxso drtax dlperinc ;

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>.289359873</td>
<td>3</td>
<td>.096453291</td>
</tr>
<tr>
<td>Residual</td>
<td>.082627329</td>
<td>44</td>
<td>.001877894</td>
</tr>
<tr>
<td>Total</td>
<td>.371987202</td>
<td>47</td>
<td>.007914621</td>
</tr>
</tbody>
</table>

Number of obs = 48
F( 3, 44) = 51.36
Prob > F = 0.0000
R-squared = 0.7779
Adj R-squared = 0.7627
Root MSE = .04333

| dlavgprs | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|---|-----|---------------------|
| drtaxso  | .013457 | .0030498 | 4.41 | 0.000 | .0073106 .0196033 |
| drtax    | .0075734 | .0010488 | 7.22 | 0.000 | .0054597 .009687 |
| dlperinc | -.0289943 | .1474923 | -0.20 | 0.845 | -.3262455 .2682568 |
| _cons    | .4919733 | .0220923 | 22.27 | 0.000 | .4474492 .5364973 |

. test drtaxso drtax;

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

F( 2, 44) = 75.65
75.65 > 10 so instruments aren’t weak
Prob > F = 0.0000

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

Copyright © 2011 Pearson Addison-Wesley. All rights reserved.
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

------------------------------------------------------------------------------

|      Coef.   Std. Err.      t P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  e |                  
  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  |   .0029390   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------

. test drtaxso drtax;
Compute J-statistic, which is m*F, where F tests whether coefficients on the instruments are zero

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

  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 $Z_{1i},...,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},...,Z_{mi}$, $W_{1i},...,W_{mi}$.

- 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
```

\[ J = 2 \times 2.47 = 4.93 \]  \quad \text{p-value from chi-squared(1) distribution}

**Now what??**
Tabular summary of these results:

<table>
<thead>
<tr>
<th>TABLE 12.1</th>
<th>Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong> $\ln(Q_{t,1995}^{\text{cigarette}}) - \ln(Q_{t,1985}^{\text{cigarette}})$</td>
<td></td>
</tr>
<tr>
<td><strong>Regressor</strong></td>
<td>(1)</td>
</tr>
<tr>
<td>$\ln(P_{t,1995}^{\text{cigarettes}}) - \ln(P_{t,1985}^{\text{cigarettes}})$</td>
<td>$-0.94^{**}$</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
</tr>
<tr>
<td>$\ln(Inc_{t,1995}) - \ln(Inc_{t,1985})$</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
</tr>
<tr>
<td>Intercept</td>
<td>$-0.12$</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
</tr>
<tr>
<td><strong>Instrumental variable(s)</strong></td>
<td>Sales tax</td>
</tr>
<tr>
<td><strong>First-stage F-statistic</strong></td>
<td>33.70</td>
</tr>
<tr>
<td><strong>Overidentifying restrictions J-test and p-value</strong></td>
<td>$-$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The J-test of overidentifying restrictions is described in Key Concept 12.6 (its $p$-value is given in parentheses), and the first-stage $F$-statistic is described in Key Concept 12.5. Individual coefficients are statistically significant at the *5% 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 (political pressure for low cigarette taxes)
  - If so, cig-only tax is endogenous
- This reasoning doesn’t apply to general sales tax

→ use just one instrument, the general sales tax
The Demand for Cigarettes: Summary of Empirical Results

- Use the estimated elasticity based on TSLS with the general sales tax as the only instrument:
  
  \[ \text{Elasticity} = -0.94, \ SE = 0.21 \]

- This elasticity is surprisingly large (not inelastic) – a 1% increase in prices reduces cigarette sales by nearly 1%. This is much more elastic than conventional wisdom in the health economics literature.

- This is a long-run (ten-year change) elasticity. What would you expect a short-run (one-year change) elasticity to be – more or less elastic?
Assess the Validity of the Study

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

External validity?

- We have estimated a long-run elasticity – can it be generalized to a short-run elasticity? Why or why not?
- Suppose we want to use the estimated elasticity of -0.94 to guide policy today. Here are two changes since the period covered by the data (1985-95) – do these changes pose a threat to external validity (generalization from 1985-95 to today)?
  - Levels of smoking today are lower than in 1985-1995
  - Cultural attitudes toward smoking have changed against smoking since 1985-95.
Where Do Valid Instruments Come From? (SW Section 12.5)

General comments
The hard part of IV analysis is finding valid instruments

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


Does cardiac catheterization improve longevity of heart attack patients?

\[ Y_i = \text{survival time (in days) of heart attack patient} \]
\[ X_i = 1 \text{ if patient receives cardiac catheterization,} \]
\[ = 0 \text{ otherwise} \]

- Clinical trials show that CardCath affects SurvivalDays.
- But is the treatment effective “in the field”?
Cardiac catheterization, ctd.

\[ \text{SurvivalDays}_i = \beta_0 + \beta_1 \text{CardCath}_i + u_i \]

- Is OLS unbiased? The decision to treat a patient by cardiac catheterization is endogenous – it is (was) made in the field by EMT technician and depends on \( u_i \) (unobserved patient health characteristics)

- If healthier patients are catheterized, then OLS has simultaneous causality bias and OLS overstates overestimates the CC effect

- Propose instrument: distance to the nearest CC hospital minus distance to the nearest “regular” hospital
Cardiac catheterization, ctd.

- $Z =$ 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 $CardCath_i$, then $\text{corr}(\text{distance}, u_i) = 0$ 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


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); benevolences = $\frac{1}{4}$ 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


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

\[ Y = 12^{\text{th}} \text{ grade test scores} \]
\[ X = \text{measure of choice among school districts (function of } \# \text{ of districts in metro area)} \]

What is the motivation for using instrumental variables?

Proposed instrument:

\[ Z = \# \text{ small streams in metro area} \]
School competition, ctd.

Data – some details

- cross-section, US, metropolitan area, late 1990s ($n = 316$),
- $Y = 12^{th}$ 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?**
Conclusion (SW Section 12.6)

• A valid instrument lets us isolate a part of $X$ that is uncorrelated with $u$, and that part can be used to estimate the effect of a change in $X$ on $Y$.

• IV regression hinges on having valid instruments:
  1. Relevance: Check via first-stage $F$.
  2. Exogeneity: Test overidentifying 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 are the threats to the internal validity of an IV regression?

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