Instrumental Variables

Chapter 12

Learning Objectives

- How to use an instrumental variable to solve common internal validity problems
- Identify key characteristics of a valid instrument and potential threats
- Test for weak instruments

Textbook Coverage

- 12.1 IV estimator with single regressor and single instrument
  - *We won't manually compute standard errors*
- 12.2 General IV regression model
- 12.3 Checking instrument validity
  - *Weak instruments and exogeneity*
  - *Exclude overidentifying restrictions test*
- 12.4/12.5 – Interesting examples!

IV Regression: Why?

Three important threats to internal validity:

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

All three problems result in $E(u|X) \neq 0$. That is, we have **endogeneity** (and violation of the zero conditional mean assumption).

Instrumental Variables Estimation and Two Stage Least Squares

- Solutions to endogeneity problems considered so far:
  - Difference in differences
  - Fixed effects models if 1) panel data is available, 2) endogeneity is time-constant, and 3) regressors are not time-constant
- Today: Instrumental variables method (IV)
  - IV is the most well-known method to address endogeneity problems
  - Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ – using an *instrumental variable* (IV), Z.

Wages and Schooling

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

- β_1 measures the returns to schooling
- One omitted variable *V*: an individuals innate ability as a worker
  - Innate ability positive affects *wages* $(\delta > 0)$
  - Likely that innate ability is positively correlated with *schooling*: corr(*education*,*V*) > 0
- Suggests OLS estimator of β_1 may have omitted variable bias
- If this is the only omitted variable, bias is positive
  - Our $\widehat{\beta_1}$ overestimates the financial returns to schooling

Wages and Schooling

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

- Data show that people who attend college earn high wages
- We want to estimate the *causal* effect
- What if we prevented someone who would like to go to college from attending college?
  - Would long-run wages be hurt by not getting schooling?

Wages and Schooling: Multiple Regression?

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

- How do we measure innate ability?
- IQ tests may measure some part of ability; hard to get IQ data for large sample
- IQ is not a perfect measure of innate ability in the workplace
  - Example: IQ test wouldn't measure social skills, which are important in the workplace
  - Note: you *should* include IQ if available
- As IQ tests are not perfect, *schooling* is likely to still be correlated with the omitted variable part of innate ability
- Then, we can't convincingly address the correlation between innate ability and schooling and include it

Wages and Schooling: Panel Data?

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

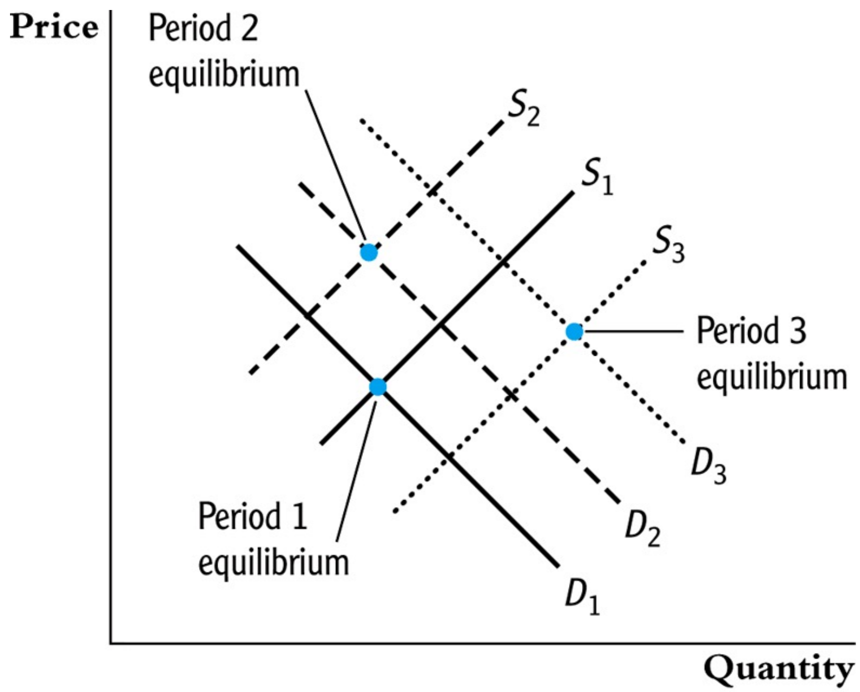
- Might be a reasonable assumption that innate ability is relatively constant over a worker's career
- But, *schooling* is also typically constant for a majority of adult workers
- Adults who go back to school after working are a non-representative group
- Panel data do not provide convincing variation in schooling over a worker's career needed to estimate the returns to schooling with worker fixed effects

Classic Example

- Estimating the demand for butter
  - Philip Wright (1928), *The Tariff on Animals and Vegetable Oils*
  - Appendix B: "The Method of Introducing External Factors": estimates the supply and demand elasticities for butter and flaxseed oil
- Wright had data on total annual butter consumption and its average annual price in the U.S. from 1912 to 1922
- Naïve estimation strategy: use OLS

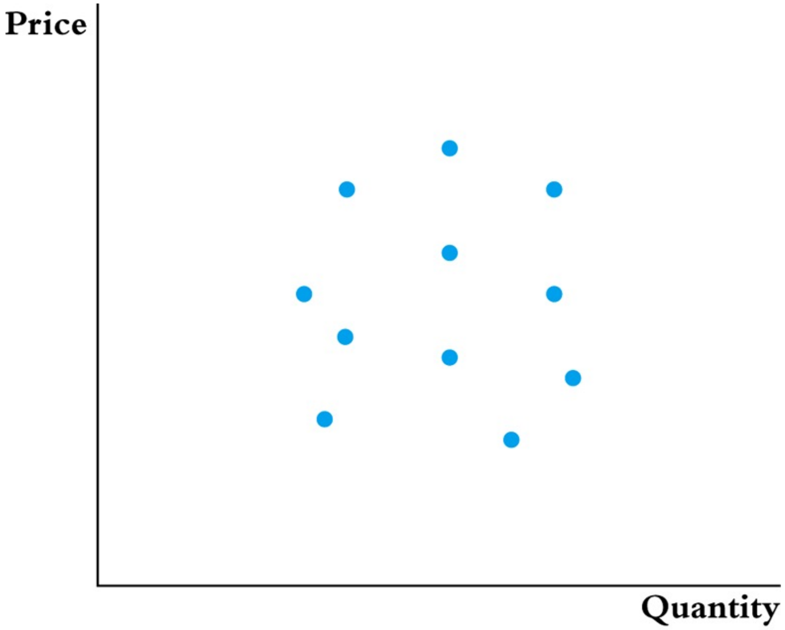
$$\ln\left(Q_i^{butter}\right) = \beta_0 + \beta_1 \ln\left(P_i^{butter}\right) + u_i$$

Reminder: Supply and Demand



(a) Demand and supply in three time periods

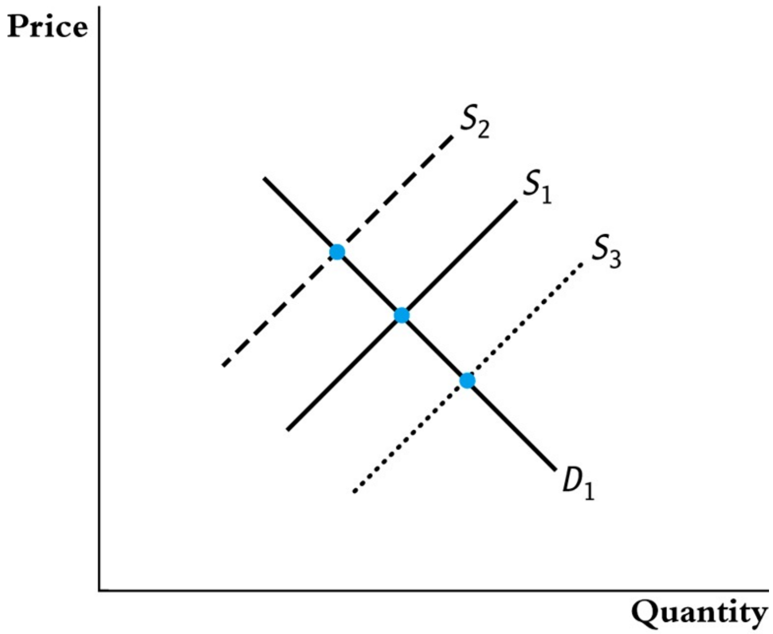
Data on Equilibrium Prices



(b) Equilibrium price and quantity for 11 time periods

- Can you tell what the supply and demand curve looks like based on these data points?

A Better Way



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

- If you can hold demand fixed, and only observe a change in supply, you can trace out the demand curve
- This is the intuition for IV

Demand for Cigarettes

- Broad public policy interest in reducing cigarette consumption
- Suppose demand for cigarettes across the 50 states:

$$sales_i = 164.4 - 0.38price_i + u_i$$

- But, price may be correlated with omitted variables in *u*
- Prices in each state determined by cigarette firms
- Cigarette firms may adjust price based on demand conditions
- When state *i* has a high u_i , this state has an unusually high demand for cigarettes
- Therefore, $price_i$ may be positively correlated with u_i

Simultaneous Causality

$$sales_i = 164.4 - 0.38price_i + u_i$$

- Simultaneous causality
  1. Y_i depends on X_i
  2. X_i depends on Y_i
- Sales depend on prices, but prices may also depend on sales
- Cigarette producers set higher prices in states where demand is stronger, where sales tend to be higher
- Simultaneous causality would *disappear* if we could randomly assign prices to the different states
  - In this experiment, there is no correlation between *price* and *u*

Simultaneous Causality

- Simultaneous causality is especially problematic because X_i will generally be correlated with *all* omitted variables in u_i
- Hard to remove omitted variable bias by measuring the omitted variables
- Would need to measure every single omitted variable

Instrumental Variables Assumptions

- An **instrumental variable** is an additional variable Z_i that satisfies three assumptions

1. Z_i **is** correlated with X_i
   - $Corr(Z, X) \neq 0$
2. Z_i **is not** correlated with the omitted variable, u_i
   - $Corr(Z, u) = 0$
3. Z_i **does not** *directly affect* (cause) Y_i
   - It can only affect Y_i through its affect on X_i
   - Z_i does not enter into the equation $Y_i = \beta_0 + \beta_1 X_i + u_i$

Identification

$$Y = \beta_0 + \beta_1 X + u$$

$$\Rightarrow$$

$$\mathrm{Cov}(Y,Z) = \mathrm{Cov}(\beta_0 + \beta_1 X + u, Z)$$

$$= \mathrm{Cov}(\beta_0, Z) + \mathrm{Cov}(\beta_1 X, Z) + \mathrm{Cov}(u, Z)$$

$$= 0 + \beta_1 \mathrm{Cov}(X,Z) + 0 \text{ by } \mathrm{Cov}(u,Z) = 0 \text{ assumption}$$

$$\Rightarrow$$

$$\mathrm{Cov}(Y,Z) = \beta_1 \mathrm{Cov}(X,Z)$$

$$\Rightarrow$$

$$\beta_1 = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)}$$

Which Assumptions Used?

$$\beta_1 = \frac{Cov(Y,Z)}{Cov(X,Z)}$$

- $Cov(Z_i, u_i) = 0$
  - explicitly used in derivation
- $Cov(X_i, Z_i) \neq 0$
  - used to divide by $Cov(X_i, Z_i)$ in solving for β_1
  - Can't divide by zero!
- Z_i does not affect Y_i directly
  - used to write down the population model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Note, we never assumed that $Cov(X_i, u_i) = 0$
  - IV explicitly allows for Omitted Variable Bias

Let's Give our Assumptions Names

1. Z_i **is** correlated with X_i
   - $\text{Corr}(Z, X) \neq 0$
   - ***Z*** is a **powerful** or **relevant** instrument
2. Z_i **is not** correlated with the omitted variable, u_i
   - $\text{Corr}(Z, u) = 0$
   - ***Z*** is an **exogenous** instrument
3. Z_i **does not** *directly affect* (cause) Y_i
   - It can only affect Y_i through its affect on X_i
   - Z_i does not enter into the equation $Y_i = \beta_0 + \beta_1 X_i + u_i$
   - ***Z*** is an **excluded** instrument

Intuition for Formula

$$\beta_1 = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)}$$

- Goal: to estimate β_1, how *X* affects *Y*
- Problem: We think *X* is correlated with *u*
- Solution: Let's not compare *Y* (which enters *u* directly) and *X* directly
  - Cov (*X* ,*Y*) explicitly not in our formula
- Instead, let's see how *Y* moves with a third variable *Z.* And, how *X* moves with *Z*
- *Z* is exogenous: uncorrelated with *u; Z* also does not affect *Y* directly
- If *Y* and *X* are both correlated with *Z*, the only explanation under our assumptions is that *X* causes *Y* according to β_1

Possible Instrument: Distance to College

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

- Schooling and ability (*V*) are correlated
- Say distance from high school to nearest college is positively correlated with schooling attainment
  - **Powerful** instrument
- And, say distance to college is uncorrelated with worker ability (*V*)
  - **Exogenous** instrument
- Assume that growing up near to a college does not *cause* your wages to be higher
  - **Excluded** instrument

Distance to College

$$\beta_1 = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)} = \frac{\mathrm{Cov}(\text{log wage,distance})}{\mathrm{Cov}(\text{schooling,distance})}$$

- Denominator is positive
- Numerator is positive if people who go to high school near to a college earn higher wages as an adult
  - Note: *not* because the distance *causes* the higher wage
- Conclude: schooling raises wages
- Returns to schooling, β_1, are high
- Cov(*X*,*Y*) does not appear in our formula
  - we do not compare someone's wage to their schooling

Two Stage Least Squares

- For a dataset with *n* observations, using sample covariance instead of population covariance
- Called **two-staged least squares**
  - Why will become apparent soon

$$\widehat{\beta_1}^{2SLS} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)} = \frac{\frac{1}{n}\sum_{i_1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n}\sum_{i_1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$\widehat{\beta_0}^{2SLS} = \bar{Y} - \widehat{\beta_1}^{2SLS}\bar{X}$$

Sales Tax and Cigarette Price

$$sales_i = \beta_0 + \beta_1 price_i + u_i$$

- Instrument for price of cigarettes?
- Need a Z_i that is
  - **Powerful**: correlated with price
  - **Exogenous:** uncorrelated with u_i (the error term for demand of cigarettes)
  - **Excluded**: does not directly impact cigarette demand
- Sales Tax in state *i*?

Sales Tax and Cigarette Price

- Sales tax in state *i* ?
- **Powerful**: Sales tax in state *i* should be positively correlated with price
  - Why? Measure price as inclusive of all sales taxes (aka what consumers pay)
- **Exogenous**: No obvious reason why sales tax should be correlated with the omitted variables u_i that determine cigarette demand
- **Excluded**: No obvious reason why sales tax would directly affect demand for cigarettes, other than through price

What are the Two Stages?

- Stage 1: A regression linking *X* and *Z*

$$X_i = \pi_o + \pi_1 Z_i + \nu_i$$

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

$$X_i = \hat{X}_i + \hat{\nu}_i$$

- Stage 2: Regress Y_i on $\widehat{X}_i$

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

Intuition for the Two Stages

- First stage regresses X on Z
- Intermediate step predicts X using Z
  - Form a best guess of X using data on Z
- We know the predicted X is not correlated with omitted variables in the second stage
  - If we predict price using sales tax, predicted prices can't be correlated with unmeasured factors that affect demand even if actual prices are
  - We assumed exogeneity: sales tax is uncorrelated with omitted variables in the second stage
- Then regress the dependent variable of interest, sales of cigarettes, on predicted prices, which are cleansed of any correlation with omitted variables
- Second stage no longer has omitted variable bias or simultaneous causality bias because we used an instrument

Stata

- Given our assumptions, 2SLS provides consistent estimates of the coefficients
- `ivregress 2sls packpc (avgprs=tax), robust`
- Dependent variable is still the first variable listed after the command 2SLS
  - `ivregress` has other options besides 2SLS
- Parenthesis before equals sign
  - Endogenous regressor
- After equals sign
  - Instrument for endogenous regressor
- Robust standard errors allow for heteroskedasticity

Stata Output

```
. ivregress 2sls packpc (avgprs=tax), robust

Instrumental variables (2SLS) regression          Number of obs   =         96
                                                  Wald chi2(1)    =      88.46
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.4219
                                                  Root MSE        =     19.567

------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avgprs |  -.4208748   .0447474    -9.41   0.000    -.5085781   -.3331715
       _cons |    169.556   7.516482    22.56   0.000      154.824    184.2881
------------------------------------------------------------------------------
Instrumented:  avgprs
Instruments:   tax
```

IV in Two Stages, Manually

```
. regress avgprs tax, robust

Linear regression                                      Number of obs =      96
                                                       F(  1,    94) =  391.06
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8089
                                                       Root MSE      =  19.289

------------------------------------------------------------------------------
             |               Robust
      avgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |   2.445839   .1236823    19.78   0.000     2.200265    2.691413
       _cons |   39.04966   5.940047     6.57   0.000     27.25556    50.84376
------------------------------------------------------------------------------

. predict avgprs_predict
(option xb assumed; fitted values)

. regress packpc avgprs_predict, robust

Linear regression                                      Number of obs =      96
                                                       F(  1,    94) =   66.21
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4123
                                                       Root MSE      =  19.938

------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
avgprs_pre~t |  -.4208748   .0517255    -8.14   0.000    -.5235769   -.3181726
       _cons |    169.556   8.168981    20.76   0.000     153.3363    185.7757
------------------------------------------------------------------------------
```

`ivregress` vs. Manual

- Two stages produce exactly the same point estimates
- However, standard errors are different
- Manual first stage has sampling errors, and Stata does not know the predicted prices used in the second stage are generated regressors
- `ivregres` command uses the correct standard error formula in the second stage
- In practice, always use `ivregres`

ivregress,first

```
. ivregress 2sls packpc (avgprs=tax), robust first

First-stage regressions
-----------------------

                                                Number of obs     =         96
                                                F(   1,      94)  =     391.06
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8089
                                                Adj R-squared     =     0.8068
                                                Root MSE          =    19.2886

------------------------------------------------------------------------------
             |               Robust
      avgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |   2.445839   .1236823    19.78   0.000     2.200265    2.691413
       _cons |   39.04966   5.940047     6.57   0.000     27.25556    50.84376
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression          Number of obs   =         96
                                                  Wald chi2(1)    =      88.46
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.4219
                                                  Root MSE        =     19.567

------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avgprs |  -.4208748   .0447474    -9.41   0.000    -.5085781   -.3331715
       _cons |    169.556   7.516482    22.56   0.000      154.824    184.2881
------------------------------------------------------------------------------
Instrumented:  avgprs
Instruments:   tax
```

Reporting the First Stage

- First stage shows how X and Z are related
- Statistical test of the assumption that X and Z are correlated
- Rule of thumb: first-stage F stat should be more than 10
  - If so, instruments are **powerful**

Weak Instruments

- What if the first-stage F test is less than 10?
- May have a “weak instrument”
- Sample covariance of X and Z may be close to 0
- Back to the definition:

$$\hat{\beta}_1^{2SLS} = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(X, Z)}$$

- Intuition: blows up your estimate

Which Assumptions Can Be Tested?

- Whether an instrument is weak or powerful can be tested by a first-stage F-test
  - If the first-stage F-test is less than 10, the standard errors reported may not have 95% coverage
- Cannot really test whether an instrument is exogenous as you lack data on the omitted variable
- Exogeneity of the instrument must be defended with reasoning about the instrument and the omitted variables in question

IV + Multiple Regression

$$sales_i = \beta_0 + \beta_1 price_i + \beta_2 income_i + u_i$$

- Measure income per person at the state level
  - Why? Income may affect sales
- Income is not determined simultaneously with the demand for cigarettes; we do not believe it is correlated with the composite omitted variable *u*
- 2SLS can handle variables not treated as endogenous, meaning not correlated with the error term

IV + Multiple Regression

```
. ivregress 2sls packpc (avgprs=tax) incomepop, robust first

First-stage regressions
-----------------------

                                                  Number of obs   =         96
                                                  F(   2,      93) =     477.80
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.9041
                                                  Adj R-squared   =     0.9020
                                                  Root MSE        =    13.7372

------------------------------------------------------------------------------
             |               Robust
      avgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   incomepop |   4.128513   .4434336     9.31   0.000     3.247941    5.009084
         tax |   1.467367   .1281227    11.45   0.000     1.212941    1.721793
       _cons |   5.821362   4.996161     1.17   0.247    -4.100024    15.74275
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression               Number of obs =      96
                                                       Wald chi2(2)  =   68.04
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.4600
                                                       Root MSE      =  18.913

------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avgprs |  -.7125521   .1355838    -5.26   0.000    -.9782915   -.4468126
   incomepop |   3.010063   1.098247     2.74   0.006     .8575387    5.162586
       _cons |   156.7195    7.32368    21.40   0.000     142.3653    171.0736
------------------------------------------------------------------------------
Instrumented:  avgprs
Instruments:   incomepop tax
```

IV + Multiple Regression

- Exogenous regressor, income per person, was added to *both* the first stage and the second stage of the regression
- Because income is assumed to be exogenous, we can use income to predict price in the first stage
- We can also use income to explain cigarette sales in the second stage
- Including income in the second stage reduces omitted variable bias with price if…
  - Income is correlated with price, and
  - Income is correlated with the instrument sales tax, so that if income was left in the omitted variable, sales tax would NOT be an exogenous instrument and 2SLS would not be consistent

Need an *Excluded* Instrument

- We need to exclude at least one instrument for each regressor treated as endogenous in the outcome equation
- Even if we have income as a regressor
- Stata will give you an error message with no excluded instrument

Panel Data, Fixed Effects, and IV

$$sales_{it} = \alpha_i + \lambda_t + \beta_1 price_{it} + \beta_2 income_{it} + \delta V_i + \omega_{it}$$

Instrument for *price* is *sales tax*

- Panel data with fixed effects can be combined with instrumental variables; data from1985 & 1995
- Include state fixed effects to control for the correlation of price and income with time-invariant omitted factors like a state's attitude towards smoking
  - Time invariant factors are in V
- Use time fixed-effects to control from correlation of price and income with factors that affect all states in one year, such as a national anti-smoking campaign
- Instrument sales tax addresses simultaneous causality between demand factors in a given state *i* and year *t*, ω_{it} , and price
- Income is again assumed to be uncorrelated with the error
- Stata command is `xtivreg`

State Fixed Effects, No Instruments

```
. egen stateID = group(state)

. xtset stateID year,yearly
       panel variable:  stateID (strongly balanced)
        time variable:  year, 1985 to 1995, but with gaps
                delta:  1 year

.
. xtreg packpc avgprs incomepop, vce(cluster state) fe

Fixed-effects (within) regression               Number of obs      =        96
Group variable: stateID                         Number of groups   =        48

R-sq:  within  = 0.9091                         Obs per group: min =         2
       between = 0.3228                                        avg =       2.0
       overall = 0.4351                                        max =         2

                                                F(2,47)            =    308.80
corr(u_i, Xb)  = 0.1321                         Prob > F           =    0.0000

                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avgprs |  -.3545608   .0578544    -6.13   0.000    -.4709489   -.2381728
   incomepop |   .2321271     .59565     0.39   0.699    -.9661661     1.43042
       _cons |   155.8269   3.350731    46.51   0.000     149.0861    162.5677
-------------+----------------------------------------------------------------
     sigma_u |  19.233659
     sigma_e |  6.1716242
         rho |  .90664991   (fraction of variance due to u_i)
```

State, Year Fixed Effects, No Instruments

```
. xtreg packpc avgprs incomepop i.year, vce(cluster state) fe

Fixed-effects (within) regression               Number of obs      =        96
Group variable: stateID                         Number of groups   =        48

R-sq:  within  = 0.9231                         Obs per group: min =         2
       between = 0.1788                                        avg =       2.0
       overall = 0.4044                                        max =         2

                                                F(3,47)            =    211.54
corr(u_i, Xb)  = -0.0032                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avgprs |  -.4159506   .0665089    -6.25   0.000    -.5497492    -.282152
   incomepop |  -1.641826   .8330128    -1.97   0.055    -3.317632    .0339791
             |
        year |
       1995  |   21.49053    6.21144     3.46   0.001     8.994728    33.98634
             |
       _cons |   187.9278    9.78275    19.21   0.000     168.2474    207.6082
-------------+----------------------------------------------------------------
     sigma_u |  19.676088
     sigma_e |  5.7389561
         rho |  .92159756   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

With Instruments, First Stage

```
. xtivreg packpc ( avgprs = taxs) incomepop y1995, vce(cluster state) fe first

First-stage within regression

Fixed-effects (within) regression               Number of obs     =         96
Group variable: stateID                         Number of groups  =         48

R-sq:                                           Obs per group:
     within  = 0.9950                                         min =          2
     between = 0.9152                                         avg =        2.0
     overall = 0.9849                                         max =          2

                                                F(3,47)           =    3001.56
corr(u_i, Xb)  = -0.0619                        Prob > F          =     0.0000

                                 (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      avgprs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   incomepop |   1.375672   .5110049     2.69   0.010     .3476629    2.403681
       y1995 |   36.03878    3.90586     9.23   0.000      28.1812    43.89636
        taxs |   1.172695   .0616058    19.04   0.000      1.04876     1.29663
       _cons |   43.76797   6.724578     6.51   0.000     30.23986    57.29607
-------------+----------------------------------------------------------------
     sigma_u |  4.5841187
     sigma_e |  4.2251491
         rho |  .54068153   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Second Stage

```
Fixed-effects (within) IV regression            Number of obs      =        96
Group variable: stateID                         Number of groups   =        48

R-sq:                                           Obs per group:
     within  = 0.9231                                          min =         2
     between = 0.1750                                          avg =       2.0
     overall = 0.4016                                          max =         2


                                                Wald chi2(3)       =   7096.08
corr(u_i, Xb)  = -0.0053                        Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      packpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     avgprs |  -.4086211   .0683084    -5.98   0.000    -.5425031   -.2747391
  incomepop |  -1.680971   .8502926    -1.98   0.048    -3.347514   -.0144281
      y1995 |   21.24298   6.233645     3.41   0.001     9.025263     33.4607
      _cons |   187.7112   9.783726    19.19   0.000     168.5355     206.887
-------------+----------------------------------------------------------------
    sigma_u |  19.722804
    sigma_e |   5.739675
        rho |  .92192154   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Instrumented:   avgprs
Instruments:    incomepop y1995 taxs
------------------------------------------------------------------------------
```

Use Logarithms

$$\log(sales_{it}) = \alpha_i + \lambda_t + \beta_1 \log(price_{it}) + \beta_2 \log(income_{it}) + \delta V_i + \omega_{it}$$

Instrument for log(price) is log(sales tax)

- Putting price in logarithms allows the time fixed effects to correct for inflation
  - Why? A dollar is worth less over time
- Correcting for inflation is also important in first stage, where price predicted using (log of) sales tax
- Coefficient on price is now elasticity of sales with respect to price, a key parameter of interest

Stata Output

```
Fixed-effects (within) IV regression            Number of obs      =         96
Group variable: stateID                         Number of groups   =         48

R-sq:                                           Obs per group:
     within  = 0.9007                                         min =          2
     between = 0.3620                                         avg =        2.0
     overall = 0.5451                                         max =          2


                                                Wald chi2(3)       =    3561.21
corr(u_i, Xb)  = 0.0363                         Prob > chi2        =     0.0000

                                  (Std. Err. adjusted for 48 clusters in state)
```

lpackpc	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lavgprs	-1.269426	.1966853	-6.45	0.000	-1.654922	-.8839299
lincomepop	.4458224	.2999498	1.49	0.137	-.1420684	1.033713
y1995	.2514037	.1901165	1.32	0.186	-.1212177	.6240251
_cons	9.508861	1.270228	7.49	0.000	7.019261	11.99846
sigma_u	.15892966					
sigma_e	.06528299					
rho	.85563024	(fraction of variance due to u_i)				

```
Instrumented:   lavgprs
Instruments:    lincomepop y1995 ltaxs
```

Comparing All Estimates

VARIABLES	(1) OLS	(2) 2SLS	(3) 2SLS	(4) State FE	(5) State and ear FE	(6) 2SLS panel	(7) 2SLS log panel
avgprs	-0.385***	-0.421***	-0.687***	-0.355***	-0.416***	-0.409***	
	(0.0412)	(0.0412)	(0.119)	(0.0579)	(0.0665)	(0.0683)	
incomepop			2.816***	0.232	-1.642*	-1.681**	
			(1.002)	(0.596)	(0.833)	(0.850)	
1995.year					21.49***	21.24***	0.251
					(6.211)	(6.234)	(0.190)
lavgprs							-1.269***
							(0.197)
lincomepop							0.446
							(0.300)
Constant	164.4***	169.6***	156.6***	155.8***	187.9***	187.7***	9.509***
	(6.700)	(7.025)	(7.256)	(3.351)	(9.783)	(9.784)	(1.270)
Observations	96	96	96	96	96	96	96
R-squared	0.426	0.422	0.463	0.909	0.923		
Number of stateID				48	48	48	48

Best Elasticity Estimate

- State fixed effects address correlation of attitudes towards smoking and cigarette prices
- Time fixed effects address say national anti-smoking campaigns that are correlated with factors affecting demand
- We add income because income is likely correlated with cigarette prices, affects sales
- Price will respond to state and time demand shocks
- Instrument for price using sales tax on cigarettes
- Our best elasticity estimate is -1.27 --> when price of cigarettes goes up by 1%, sales go down by 1.3%
    - Point estimate shows that demand is elastic, but not terribly so
- However, confidence interval of (-1.53,-1.004) barely excludes -1, so we can statistically reject the null hypothesis that demand for cigarettes is inelastic

Example #1: Effect of Studying on Grades

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

Y = GPA

X = study time (hours per day)

Data: grades and study hours of college freshmen.

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

Studying on grades, ctd.

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

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

Studying on grades, ctd.

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

1. Is the instrument **powerful**?
2. Is the instrument **exogenous**?
3. Is the instrument **excludable**?

Evidence

Table 2
First Stage Regressions
The effect of instruments (and other variables) on study hours

Independent Variable	estimate (std error) n=210	estimate (std error) n=176
INSTRUMENTS		
video game TREATMENT	-.668 (.252)**	-.658 (.268)**
RSTUDYHS		.028 (.013)**
REXSTUDY		.049 (.074)

Table 4
Estimates of the effect of studying on grade performance:
Ordinary Least Squares, Instrumental Variables, Fixed Effects

Independent Variable	OLS n=210 estimate (std. error)	IV instrument: video game TREATMENT n=210 estimate (std. error)	IV instruments: video game TREATMENT, RSTUDYHS, REXSTUDY n=176 estimate (std. error)
CONSTANT	.719 (.408)*	-.073 (.709)	-.062 (.638)
STUDY	.038 (.025)	.360 (.183)**	.291 (.121)**
SEX	-.132 (.084)	-.023 (.129)	-.010 (.126)

Returns to Schooling

$$\log(wage_i) = \beta_0 + \beta_1 schooling_i + \delta V_i + u_i$$

- Data show that people who attend college earn high wages
- We want to estimate the *causal* effect
- OLS isn't able to distinguish whether high wages are due to the causal benefit of schooling or because people who attend college would be able workers no matter what their schooling level
  - Innate ability in *u*
- At the extreme: college might just be a way to signal to employers a student's innate ability; credentials how innately smart you are

Instrument: Quarter of Birth

- Many states/school districts do not let you drop out until age 16 (some places 17)
- High school students turn age 16 at different times during the year
- Children born earlier in the year can drop out earlier
- So, children born earlier in the year get less total schooling
- Angrist and Krueger (1991)

IV Assumptions

- **Relevance (power)** – can test this empirically, but cannot shift total schooling more than a few months
- **Exogenous –** Unlikely your innate ability as a worker is correlated with your quarter of birth
- **Exclusion** – Unlikely quarter of birth directly affects your wages

Quarter of Birth Effects: First Stage

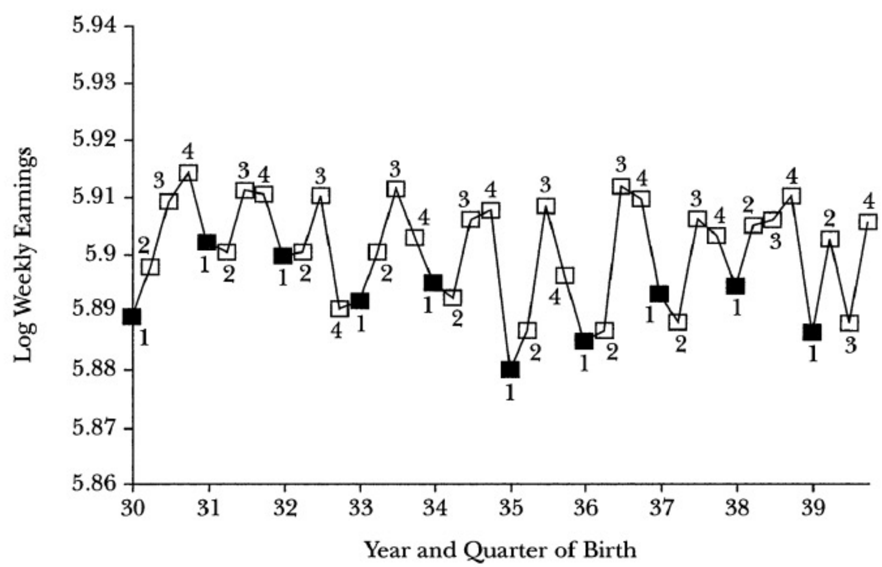
Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]

Second Stage

- Dependent Variable – log of wage
- Regressor of interest – years of schooling
- Instruments: Quarter of Birth dummies, interacted with year of birth dummies
- Non-endogenous regressors – year of birth, other covariates shown in coming tables

Weekly Earnings by Quarter of Birth

Figure 2
Mean Log Weekly Earnings, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

Returns to Schooling

TABLE IV
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

a. Standard errors are in parentheses. Sample size is 247,199. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the State, County, and Neighborhoods 1 percent samples of the 1970 Census (15 percent form). The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

Returns to Schooling

- OLS and 2SLS estimates *quite similar* for all specifications
- 2SLS standard errors are higher
- Putting in age and age squared makes 2SLS higher than OLS
- Cannot statistically reject 2SLS different than OLS in any specification

Weak Instruments?

- Might have a weak instrument
  - Sample covariance of Z and X may be near to 0
  - Dividing by a number close to 0 in

$$\beta_1 = \frac{Cov(Y,Z)}{Cov(X,Z)}$$

Maybe Cov(Z, *u*) is not 0

- Unlikely quarter of birth is completely unrelated to innate ability and other factors
- Unlikely quarter of birth directly excludable from outcome equation
- Bound, Jaeger, Baker (1993) cite references
- Quarter of Birth related to
  - School attendance
  - Behavioral difficulties by students
  - Mental health referrals
  - Performance in reading, writing, arithmetic
  - Schizophrenia
  - IQ
  - Family Incomes

Low First-Stage F-Stats

Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings
(standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.029)
F (excluded instruments)		13.486		4.747		1.613
Partial R^2 (excluded instruments, ×100)		.012		.043		.014
F (overidentification)		.932		.775		.725
Age Control Variables						
Age, Age^2	x	x			x	x
9 Year of birth dummies			x	x	x	x
Excluded Instruments						
Quarter of birth		x		x		x
Quarter of birth × year of birth				x		x
Number of excluded instruments		3		30		28

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. F (first stage) and partial R^2 are for the instruments in the first stage of IV estimation. F (overidentification) is that suggested by Basmann (1960).

Weak Instruments?

- Rule of thumb: F-stat of 10 or greater on the *excluded* instruments
- With proper age controls as additional regressors in first and second stages, Bound et al find an *F*-stat of 1.6
- Angrist and Krueger's regressions had a weak instrument
- Combined with a small correlation of the excluded instruments with *u*, a weak instrument could result in important bias in the estimates of returns to schooling
- Theory in Bound et al suggests weak instruments should lead IV estimates to look the same as OLS

Conclusion?

- Instruments are a *powerful* tool in econometrics
- With the right instrument you can get a quasi-experimental design and causal estimates
- With the wrong estimate you can introduce quite a bit of bias in your regressions
- There are some guidance metrics (F-stat), but coming up with an instrument relies on a lot of background knowledge, and sometimes luck