ECON 7310 Elements of Econometrics
Week 10: Instrumental Variables Regression

David Du[1]

[1] University of Queensland

Draft

Outline

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

IV Regression: Why?

- Three important threats to internal validity are:
  - Omitted variable bias from a variable that is correlated with X but is unobserved (so cannot be included in the regression) and for which there's no adequate control variable;
  - Simultaneous causality bias (X causes Y, Y causes X);
  - Errors-in-variables bias (X is measured with error)
- All three problems result in $E(u|X) \neq 0$.
- Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ by using an instrumental variable (IV), Z.

The IV Estimator with one Regressor and one Instrument SW Section 12.1

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

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an instrumental variable, Z_i, which is correlated with X_i but uncorrelated with u_i.
- By exploiting the correlation of Z_i and X_i, we obtain a consistent estimator.

Endogeneity and Exogeneity, and Conditions for a Valid Instrument

- Endogeneity and Exogeneity
  - An **endogenous** variable is one that is correlated with u
  - An **exogenous** variable is one that is uncorrelated with u
- For an instrumental variable (an **instrument**) Z to be valid, it must satisfy two conditions:
  1. **Instrument relevance:** $corr(Z_i, X_i) \neq 0$
  2. **Instrument exogeneity:** $corr(Z_i, u_i) = 0$
- Suppose for now that you have such a Z_i (we will discuss how to find instrumental variables later).
- How to use Z_i to estimate β_1?

The IV estimator with one X and one Z

Two Stage Least Squares (TSLS):
As it sounds, TSLS has two stages – two regressions:

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

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

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i.
- (π_0, π_1) unknown. So, we use consistent estimates $(\widehat{\pi}_0, \widehat{\pi}_1)$, i.e., OLS.
- Compute the predicted values of X_i,

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

for $i = 1, \ldots, n$.

Two Stage Least Squares (continued)

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

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

- Because $\widehat{X}_i$ is not correlated with u_i, the first least squares assumption, $E[u|\widehat{X}] = 0$ holds here (when n is large).
- Thus, β_1 can be estimated by regressing Y on $\widehat{X}$ by OLS
- The resulting estimator is called the Two Stage Least Squares (TSLS) estimator, $\widehat{\beta}_1^{TSLS}$.

How does IV work?

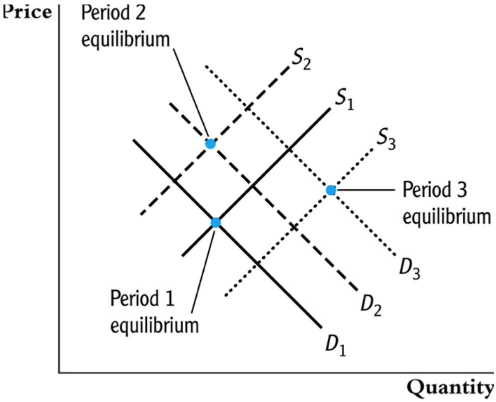
Example # Philip Wright's problem:

- Philip Wright was concerned with an important economic problem of his day (1920s): how to set an import tariff such as butter.
- Observe data on butter quantity Q_i and price P_i each year (US).
- The key is to estimate demand and supply elasticities. So, log-log form

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$$

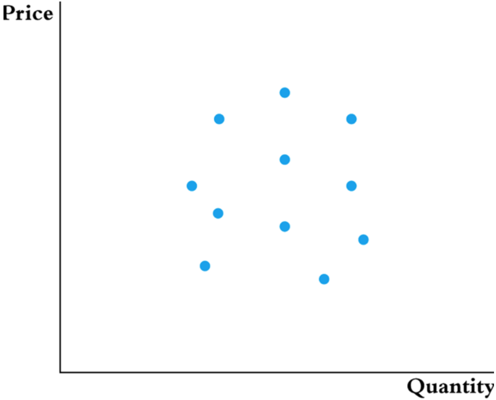
- If we run OLS, is $\widehat{\beta}_1$ the price elasticity of demand? or supply?
- In fact $\widehat{\beta}_1$ suffers from simultaneous causality bias because price and quantity are determined by the interaction of demand and supply:

simultaneous causality bias in supply and demand



(a) Demand and supply in three time periods

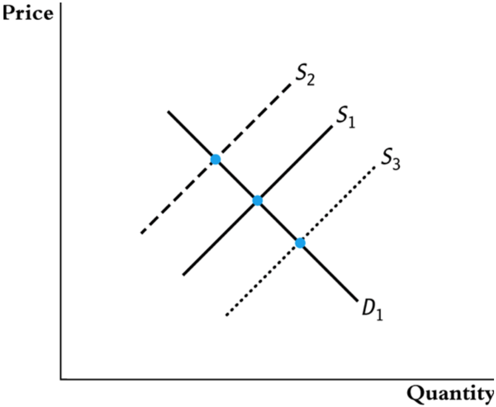
data scatter diagram must look like



(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?

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



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

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

TSLS in the supply-demand example:

- Regression equation: $\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$
- Let Z = rainfall in dairy-producing regions. Is Z a valid instrument?
  1. **Instrument relevance:** $corr(Z_i, \ln(P_i)) \neq 0$?
     Plausibly: insufficient rainfall $\Rightarrow$ less grazing $\Rightarrow$ butter supply $\downarrow$ $\Rightarrow$ prices $\uparrow$
  2. **Instrument exogeneity:** $corr(Z_i, u_i) = 0$?
     Plausibly: rainfalls in Europe does not *directly* affect demand for butter in US
- Two Stage Least Squares:

Stage 1: Regress $\ln(P_i)$ on Z_i, compute fitted value $\widehat{\ln(P_i)}$
$\Rightarrow$ isolates part of $\ln(P_i)$ that is explained by supply shifts (not by demand)

Stage 2: Regress $\ln(Q_i)$ on $\widehat{\ln(P_i)}$, compute fitted value
$\Rightarrow$ uses shifts in the supply curve to trace out the demand curve

Statistical Properties of $\widehat{\beta}_1^{TSLS}$

- $\widehat{\beta}_1^{TSLS}$ is **consistent** ($\widehat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1$) and **asymptotically normal**.
- By asymptotic normality, we can conduct hypothesis testing and construct confidence intervals.
- Note that the OLS standard error in Stage 2 is misleading because it does not take into account the fact that the regressor is a fitted value $\widehat{X}$.
- Most econometric softwares automatically computes correct $SE(\widehat{\beta}_1^{TSLS})$.
- Then, a 95% confidence interval is given by

$$\widehat{\beta}_1^{TSLS} \pm 1.96 SE(\widehat{\beta}_1^{TSLS})$$

Application: Demand for Cigarettes

- US government wishes to impose tax on cigarettes to reduce cigarette consumption $\Rightarrow$ to reduce illnesses and deaths from smoking, social costs, negative externalities, etc.
- So, it is critical to know the price elasticity of cigarette demand.
  - Suppose it is aimed to reduce cigarette consumption by 20%.
  - If the price elasticity is -0.5, the price has to increase by 40%.
- So, we consider a log-log specification.

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$$

  where Q_i is annual cigarette consumption P_i is average price including tax for state $i = 1, \ldots, 48$. (In fact, panel data 1985-1995)
- Supply-Demand interact $\Rightarrow$ OLS will suffer simultaneity bias.

Application: Demand for Cigarettes, continued

- Again, the regression equation is

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + u_i$$

- Proposed IV: Z_i = general sales tax per pack = $SalesTax_i$
  - **Instrument relevance:** $corr(SalesTax_i, \ln(P_i)) \neq 0$
  - **Instrument exogeneity:** $corr(SalesTax_i, u_i) = 0$
- Relevance should be fine because $SalesTax_i \uparrow \Rightarrow P_i \uparrow$
- Exogeneity: $SalesTax_i$ affects $\ln(Q_i)$ only indirectly through $\ln(P_i)$
  - Each state i chooses $SalesTax_i$ depending on a number of elements such as income tax, property tax, other taxes to finance its public undertakings
  - Those choices about public finance are driven by political considerations, not by demand for cigarettes.
  - So, it is plausible that $SaleTax_i$ is exogenous.

Application: Demand for Cigarettes, Stage 1

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

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

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

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

Application: Demand for Cigarettes, Stage 2

```
          Y    X-hat
. reg lpackpc lravphat if year==1995, r;

Regression with robust standard errors                 Number of obs =      48
                                                       F(  1,    46) =   10.54
                                                       Prob > F      =  0.0022
                                                       R-squared     =  0.1525
                                                       Root MSE      =  .22645

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

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

Application: Demand for Cigarettes, All at once

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

Instrumental variables (2SLS) regression               Number of obs =      48
                                                       Wald chi2(1)  =   12.05
                                                       Prob > chi2   =  0.0005
                                                       R-squared     =  0.4011
                                                       Root MSE      =  .18635

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3122035    -3.47   0.001    -1.695494   -.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 varible
------------------------------------------------------------------------------
```

Estimated cigarette demand equation:

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

Summary So far

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

The General IV Regression Model SW Section 12.2

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors ($X_1, \ldots, X_k$)
  - multiple included exogenous variables ($W_1, \ldots, W_r$) 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 $\widehat{X}$.
- 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 linear regression problems, identification depends on the number of instruments (m) and the number of endogenous regressors (k).

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

  - $X_{1i}, \ldots, X_{ki}$: **endogenous regressors** (potentially correlated with u_i)
  - $W_{1i}, \ldots, W_{ri}$: **included exogenous regressors** (uncorrelated with u_i)
  - $Z_{1i}, \ldots, Z_{mi}$: **instrumental variables** (**excluded exogenous variables**)
- $\beta_1, \ldots, \beta_k$ are said to be
  - **exactly identified** if $m = k$, e.g. we studied so far $k = 1$ and $m = 1$.
  - **overidentified** if $m > k$
  - **underidentified** if $m < k$, e.g., if $k = 1$ but $m = 0$, no identification!

TSLS with a Single Endogenous Regressor

- Consider the regression model;

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

- We have m instruments: $Z_1, \ldots, Z_m$.

Stage 1: Regress X on all the exogenous regressors $(W_1, \ldots, W_r)$ and $(Z_1, \ldots, Z_m)$, and an intercept, by OLS. Obtain predicted values $\widehat{X}$

Stage 2: Regress Y on $\widehat{X}$, $(W_1, \ldots, W_r)$, and an intercept, by OLS

- The coefficients from this second stage regression are the TSLS estimators, but SEs are wrong
- To get correct SEs, do this (in a single step) using your regression software

Demand for cigarettes, continued

- We will estimate the regression model

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + \ln(Income_i) + u_i$$

- We will use two ($m = 2$) instruments: general sales tax (Z_1) and cigarette specific tax (Z_2).
- Suppose income is exogenous (this is plausible ? why?), and we also want to estimate the income elasticity:
- Endogenous variable: $\ln(P_i)$, so $k = 1$
- Since ($m > k$), β_1 is over-identified.

Example: Cigarette demand, one instrument

```
IV: rtaxso = real overall sales tax in state
                 Y      W        X          Z
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r;

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

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495    -.413375     .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112    11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso        STATA lists ALL the exogenous regressors
                                     as instruments – slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

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

Example: Cigarette demand, two instruments

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

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

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

- Smaller SEs for $m = 2$. Using 2 instruments gives more information
- Low income elasticity (not a luxury good), though insignificantly
- Surprisingly high price elasticity

TSLS with Multiple Endogenous Regressors

Idea is exactly the same as the case with $k = 1$. Just apply Step 1 for all endogenous variables.

- Consider the regression model;

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

- We have m instruments: $Z_1, \ldots, Z_m$ with $m \geq k$.

Stage 1: Regress <u>each</u> of $X_1, \ldots, X_k$ on all the exogenous regressors $(W_1, \ldots, W_r)$ and $(Z_1, \ldots, Z_m)$, and an intercept, by OLS. Obtain predicted values $\widehat{X}_1 \ldots, \widehat{X}_k$

Stage 2: Regress Y on $\widehat{X}_1 \ldots, \widehat{X}_k$, $(W_1, \ldots, W_r)$, and an intercept, by OLS

- To get correct SEs, do this (in a single step) using your regression software

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 relevance:**
   - General case, multiple X's: Suppose the second stage regression could be run using the predicted values from the population first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression.
   - Special case of one X: the general assumption is equivalent to (a) at least one instrument must enter the population first stage regression, and (b) the W's are not perfectly multicollinear.
2. **Instrument exogeneity:** $corr(Z_{1i}, u_i) = 0, \ldots, corr(Z_{mi}, u_i) = 0$

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. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$. That is, Ws are really exogenous.
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ are i.i.d.
3. The X's, W's, Z's, and Y have nonzero, finite 4th moments
4. The instruments $(Z_{1i}, \ldots, Z_{mi})$ are <u>valid</u>.

- Under 1-4, TSLS and its t-statistic are normally distributed

Checking Instrument Validity (SW Section 12.3)

Recall the two requirements for valid instruments:

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

- What happens if one of these requirements is not satisfied? How can you check? What do you do?
- If you have multiple instruments, which should you use?

Checking 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 **weak** if $\pi_1, \ldots, \pi_m$ are all either zero or nearly zero.
- When the instruments are weak, the usual methods for statistical inference are misleading even if n is large.

Checking Instrument Relevance

- First stage regression:

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

- We consider the hypothesis that all instruments are not relevant, i.e., $\pi_1 = \cdots = \pi_m = 0$
- Rule of Thumb:
  - Compute F-statistic for $H_0 : \pi_1 = \cdots = \pi_m = 0$
  - We do not worry about weak instruments if the **first stage F statistic** >10.
  - Why 10? See Appendix 12.5.
- What do we do if instruments are weak?
  - When overidentified ($m > k$), discard weak instruments.
  - When $m = k$, find stronger instruments (or use a correct inference procedure, but this is beyond scope of the course!)

Checking Instrument Exogeneity

1. **Case of exact-identification** ($m = k$): there is no way to statistically test the assumption of instrument exogeneity.
   - necessary to use expert judgment based on personal knowledge
2. **Case of over-identification** ($m > k$):
   - There is no way to statistically test instrument exogeneity for all instruments
   - But, if some of instruments are certainly exogenous, we can test exogeneity of the other instruments.
   - This test is called the **overidentifying restrictions test**.

- Idea of overidentifying restrictions test: ($k = 1$ and $m = 2$)
  - Z_1 is exogenous for sure and want to test Z_2.
  - Suppose that $\widehat{\beta}^{TSLS}$ uses only Z_1 and $\widetilde{\beta}^{TSLS}$ uses only Z_2.
  - We know $\widehat{\beta}^{TSLS} \xrightarrow{p} \beta$ for sure. If Z_2 is exogenous, it should be $\widetilde{\beta}^{TSLS} \xrightarrow{p} \beta$
  - So, $\widetilde{\beta}^{TSLS}$ is very different from $\widehat{\beta}^{TSLS}$, it is evidence against exogeneity of Z_2.

Overidentifying Restrictions Test (The J-Statistic)

- Overidentifying restrictions test carries out this idea implicitly. Ideally, want to test $corr(u, Z) = 0$, but u is unobservable. So, we use

$$\widehat{u}^{TSLS} := Y_i - (\widehat{\beta}_0^{TSLS} + \widehat{\beta}_1^{TSLS} X_1 + \cdots + \widehat{\beta}_k^{TSLS} X_k + \widehat{\beta}_{k+1}^{TSLS} W_1 + \cdots + \widehat{\beta}_{k+r}^{TSLS} W_r)$$

  where we use the original regressors (X) not the predicted ones ($\widehat{X}$)
- Test procedure (choose a significance level α first):
  - Use OLS to estimate the coefficients in

$$\widehat{u}^{TSLS} = \delta_0 + \delta_1 Z_1 + \cdots + \delta_m Z_m + \delta_{m+1} W_1 + \cdots + \delta_{m+r} W_r + e$$

  - If $corr(Z_j, u) = 0$ for all $j = 1, \ldots, m$, we must have $\delta_1 = \cdots = \delta_m = 0$
  - Compute homoskedasticity-only F-statistic testing $H_0 : \delta_1 = \cdots = \delta_m = 0$.
  - Then, compute the J statistic $J := mF \sim \chi^2_{m-k}$.
  - Reject H_0 if $J >$ critical value at α: see the prob table of χ^2_{m-k}
    Or, reject H_0 if p-value < your significance level α.

Overidentifying Restrictions Test (The J-Statistic)

$$J := mF \sim \chi^2_{m-k}$$

- Here, $m - k$ is the degree of freedom = #. of over-identifying restrictions.
- Rejecting $H_0 \Rightarrow$ we have statistical evidence against H_0 at the chosen α. So, at least one of Zs may not be exogenous.
- The J statistic for Heteroskedastic errors is given in SW Section 19.7.
- When $m = k$, $J = 0$, always!
- This makes sense: there is no way to test exogeneity of instruments if exactly identified.

Application: Demand for Cigarettes (SW Section 12.4)

- Why are we interested in knowing the elasticity of demand for cigarettes?
- Theory of optimal taxation.
  - optimal tax rate $\propto$ 1/price elasticity
  - if demand is highly sensitive to price change, the tax rate should be small.
- Negative externalities – the government should intervene to reduce smoking
  - health effects of second-hand smoke? (non-monetary)
  - monetary externalities
- Panel Data on 48 US states (1985-1995): annual cigarette consumption, average prices, income, tax rates (cigarette specific, general commodity)

Fixed Effects model of cigarette demand

Regression model:

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

where $i = 1, \ldots, 48$ and $t = 1985, \ldots, 1995$

- State FE, α_i, reflects unobserved omitted factors that vary across states but not over time, e.g. attitude towards smoking
- Even after controlling for the FE, $corr(\ln(P_{it}), u_{it})$ is plausibly nonzero because of supply/demand interactions
- So, use TSLS to handle simultaneous causality bias
- However, the demand for addictive products like cigarettes might be inelastic in the short run. That is, the short-run elasticity ≈ 0.

The "Change" Method, $T = 2$

- So, we use $T = 2$ only with 1985 and 1995 ("changes" method) to focus on the long-term response, not short-term dynamics
- Regression equations for $t = 1985$ and 1995;

$$\ln(Q_{i,85}) = \alpha_i + \beta_1 \ln(P_{i,85}) + \beta_2 \ln(Income_{i,85}) + u_{i,85}$$
$$\ln(Q_{i,95}) = \alpha_i + \beta_1 \ln(P_{i,95}) + \beta_2 \ln(Income_{i,95}) + u_{i,95}$$

- Difference:

$$\begin{aligned}[\ln(Q_{i,95}) - \ln(Q_{i,85})] =& \beta_1 [\ln(P_{i,95}) - \ln(P_{i,85})] \\ & \beta_2 [\ln(Income_{i,95}) - \ln(Income_{i,85})] + (u_{i,95} - u_{i,85})\end{aligned}$$

- Equivalently,

$$\ln\left(\frac{Q_{i,95}}{Q_{i,85}}\right) = \beta_1 \ln\left(\frac{P_{i,95}}{P_{i,85}}\right) + \beta_2 \ln\left(\frac{Income_{i,95}}{Income_{i,85}}\right) + e_i$$

where $e_i := u_{i,95} - u_{i,85}$.

Stata: Cigarette Demand

- First, define variables;

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

One instrument, Z_1 = general sales tax only

```
                        Y        W          X          Z
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso) , r;

IV (2SLS) regression with robust standard errors       Number of obs =      48
                                                       F(  2,    45) =   12.31
                                                       Prob > F      =  0.0001
                                                       R-squared     =  0.5499
                                                       Root MSE      =  .09092

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

NOTE:

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

Instrument relevance: First Stage F statistic > 10 ?

```
.  reg dlavgprs drtaxso dlperinc;

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  2,    45) =   23.86
       Model |  .191437213     2  .095718606           Prob > F      =  0.0000
    Residual |  .180549989    45  .004012222           R-squared     =  0.5146
-------------+------------------------------           Adj R-squared =  0.4931
       Total |  .371987202    47  .007914621           Root MSE      =  .06334

------------------------------------------------------------------------------
    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
       F(  1,    45) =   46.41              the square of the t-stat:
            Prob > F =    0.0000            6.81*6.81 = 46.41

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

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

Two Instruments, adding Z_2 = cigarette specific tax only

```
                     Y          W         X          Z1     Z2
. ivregress 2sls dlpackpc dlperinc (dlavgprs = drtaxso drtax) , vce(r);

Instrumental variables (2SLS) regression               Number of obs =       48
                                                       Wald chi2(2)  =    45.44
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =   0.5466
                                                       Root MSE      =   .08836


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


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

First-stage F – both instruments

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

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =   51.36
       Model |  .289359873     3  .096453291           Prob > F      =  0.0000
    Residual |  .082627329    44  .001877894           R-squared     =  0.7779
-------------+------------------------------           Adj R-squared =  0.7627
       Total |  .371987202    47  .007914621           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…

Test the overidentifying restrictions

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

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

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     drtaxso |   .0127669   .0061587     2.07   0.044      .000355    .0251789
       drtax |  -.0038077   .0021179    -1.80   0.079     -.008076    .0004607
    dlperinc |  -.0934062   .2978459    -0.31   0.755    -.6936752    .5068627
       _cons |    .002939   .0446131     0.07   0.948    -.0869728    .0928509
------------------------------------------------------------------------------
.  test drtaxso drtax;
 ( 1)  drtaxso = 0                   Compute J-statistic, which is m*F,
 ( 2)  drtax = 0                     where F tests whether coefficients on
                                     the instruments are zero
       F(  2,    44) =    2.47       so J = 2 × 2.47 = 4.93
            Prob > F =    0.0966           ** WARNING - this uses the wrong d.f. **
```

Test the overidentifying restrictions

- Recall that $J = m \times F = 2 \times 2.47 = 4.94$. which is distributed as χ^2_{2-1} if both instruments are exogenous H_0
- The critical value at 5% level is 3.84 (see the prob table of χ^2 distributions)
- Hence, we reject $H_0 \Rightarrow$ at least one of the instruments is not exogenous. The J-test doesn't tell us which! You must exercise judgment...
- Z_2 (cig-only tax) can be endogenous, e.g., lots of smokers (high u) could have political power to keep Z_2 at a low level.

Estimation Results

TABLE 12.1 Two Stage Least Squares Estimates of the Demand for Cigarettes Using Panel Data for 48 U.S. States

Dependent variable: $\ln(Q^{cigarettes}_{i,1995}) - \ln(Q^{cigarettes}_{i,1985})$

Regressor	(1)	(2)	(3)
$\ln(P^{cigarettes}_{i,1995}) - \ln(P^{cigarettes}_{i,1985})$	−0.94** (0.21)	−1.34** (0.23)	−1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage F-statistic	33.70	107.20	88.60
Overidentifying restrictions J-test and p-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 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.

- Elasticity=0.94: a 1% increase in prices $\downarrow\downarrow$ cigarette sales by 0.94%.
- Increased taxes can substantially discourage cigarette consumption, at least in the long run

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

The hard part of IV analysis is finding valid instruments

- **Method 1: economic theory**
  - Find a variable Z that shifts only the supply curve. Then, Z is an IV for estimation of demand.
  - For example, rainfalls in Europe would changes butter production but don't change demand for butter in US
- **Method 2: exogenous source of variation in X**
  - look for exogenous variation (Z) that is "as if" randomly assigned (does not directly affect Y) but affects X
  - This approach requires knowledge of the problem being studied and careful attention to the details of data
  - Some examples follow

Example 1: Does putting criminals in jail reduce crime?

- Answer should be 'YES', but question is how much? How much the crime rate would decrease when the prison population increases by 1%?
- Variables in regression analysis using state data, e.g., $i = 1, \ldots, 48$.
  - Y_i : crime rate
  - X_i : incarceration rate, β_1
  - W_i : control variables (economic conditions and demographics)
- Estimating β_1 by OLS might suffer simultaneity bias. i.e., Y causes X
  - the simultaneity bias cannot be solved by better controls.
  - but a good instrument can fix this problem
- Potential instrument Z: prison capacity for each i
  - Relevance: small $Z \rightarrow$ release criminals $\rightarrow$ large X, so $corr(Z, X) \neq 0$.
  - Exogeneity: Z would not directly affect Y, so $corr(Z, u) = 0$.

Example 2: Does aggressive treatment of heart attacks prolong lives?

- Variables in regression analysis, patients are indexed by $i = 1, \ldots, n$.
  - Y_i : survival time (days) after heart attack
  - X_i : dummy for cardiac catheterization, β_1 (putting a tube into a blood vessel)
  - W_i : control variables (age, weight, other variables), correlated with mortality
- OLS estimate for β_1 suffers bias: $X_i = 1$ is a decision of the patient & doctor in part based on unobserved factors. So, $corr(X_i, u_i) \neq 0$.
- A potential instrument Z: distance from patient i's home to the nearest cardiac catheterization hospital
  - Relevance: smaller $Z \to$ easier to get treatment $X = 1$, so $corr(Z, X) \neq 0$.
  - Exogeneity: Z would not directly affect Y, so $corr(Z, u) = 0$.

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:
  - Relevance: Check via first-stage F, rule of thumb $F > 10$
  - 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.