Lecture notes to Stock and Watson chapter 12

Instrument variable regression

Tore Schweder

October 2009

Outline

- Do SW: 11.6
- Exogenous and endogenous regressors
- The problem of estimating a demand function
- Untangling the endogeneity by instrument variables
- Relevance and validity of instrument variables
- TSLS: Two Stage Least Squares
- Several endogenous regressors and several instruments: over- exact- and under-identification
- Testing for relevance and validity of instruments
- Weak instruments make non-sense
- On the art of finding good instruments, and of arguing their value

Example: how much tax on petrol to reduce consumption by 25%?

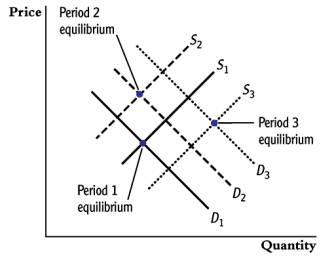
Suppose petrol consumption must be reduced by 25% due to climate considerations. How high must the price per liter be to achieve the goal? Data: Price P and quantity Q in various markets (by region and period).

- We want the demand function $E\left(Q\middle|\text{ do } P=p\right)=q(p)$, i.e. what mean consumption would have been were price determined (through tax on petrol, or by dictum) to p
  - Conditional distributions etc. by the "do" operator id due to Judea Pearl (2001) Causality: Models, Reasoning, and Inference. Cambridge University Press
  - The standard conditional expectation $E\left(Q\middle|\ P=p\right)$ represent association: mean quantity in cases P happen to be p.
    - The interaction between demand and supply make Q and P to be simultaneously determined - P is endogenous, and is correlated with Q both through the demand and the supply curves.
  - Regression of Q on P gives us $E\left(Q\middle|\ P=p\right)$ but not $E\left(Q\middle|\text{ do } P=p\right)=q(p)$.

Example cont.

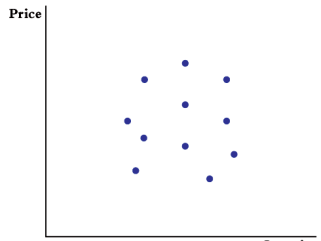
FIGURE 12.1

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



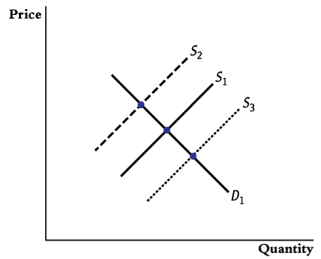
(a) Demand and supply in three time periods

(b) This scatterplot shows equilibrium price and quantity in 11 different time periods. The demand and supply curves are hidden. Can you determine the demand and supply curves from the points on the scatterplot?



(b) Equilibrium price and quantity for 11 time periods

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



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

Figure: A naive regression of quantum on price might be grossly misleading for $q(p)$!

Example cont.

- In our market causality flows both ways $Q \leftrightarrow P$
  - In purely quantum-regulated markets, $Q \longrightarrow P$. Data from strictly quantum-regulated markets are unusable.
  - Strict price regulation: $Q \longleftarrow P$. Data directly usable to estimate $q(p)$.
- Way out when causality flows both ways: find an instrument variable Z which is correlated with P, but not with the residual $Q - q(p)$.
  - Possible instruments
    1. number cars on the road in China
    2. tension between Iran and USA, difficulties in passing the Hormouz strait
    3. stop in oil pipes from production sites

Exogenous and endogenous regressors

Desired. $\mu(x) = E(Y|\text{ do } X = x) = \beta_0 + \beta_1 x$

- X is **endogenous** when causality flows both ways:
  $cov(X, Y - \mu(X)) = cov(X, u) \neq 0 \quad u = Y - (\beta_0 + \beta_1 x)$
- X is **exogenous** when $X \longrightarrow Y$
- Z is a useful instrument variable when
  1. Z is **exogenous**: $cov(Z, u) = 0$
     - Z is only correlated with Y through X
     - Z is exogenous to the Y and X interaction
  2. Z is **relevant**: $cov(Z, X) \neq 0$
     - Z is a predictor for X
- Z is **valid** when the flow of causality is $Z \longrightarrow X \longleftrightarrow Y$

TSLS

Two stage least squares for one instrument and one endogenous regressor

1. The linear regression of X on Z:
   $E(X|Z) = \pi_0 + \pi_1 Z, \quad \pi_1 = \frac{cov(X,Z)}{var(Z)}$
2. The linear regression of Y on Z:
   $E(Y|Z) = E(E(Y|\text{ do } X, Z)|Z) = E(\beta_0 + \beta_1 X|Z) =$
   $\beta_0 + \beta_1 (X|Z) = \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 Z = \gamma_0 + \gamma_1 Z$
   - Since $E(Y|Z) = \beta_0 + \beta_1 \pi_0 + \frac{cov(Y,Z)}{var(Z)} Z, \quad \beta_1 = \frac{\gamma_1}{\pi_1} = \frac{cov(Y,Z)}{cov(X,Z)}$

- Empirically:
$$\widehat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$
- Since both $s_{YZ} \xrightarrow{P} cov(Y, Z)$ and $s_{XZ} \xrightarrow{P} cov(X, Z)$, $\widehat{\beta}_1^{TSLS} \xrightarrow{P} \beta_1$: The estimator is consistent and asymptotically normal (Z valid and relevant).

The large-sample variance is

$$var\left(\widehat{\beta}_1^{TSLS}\right) \approx \frac{1}{n}\frac{var\left(\left(Z-\mu_Z\right)u\right)}{cov\left(Z,X\right)^2} = \frac{1}{n}\frac{var\left(Z-\mu_Z\right)var(u)}{cov\left(Z,X\right)^2}$$

when Z and u are independent (not only un-correlated):

$$\begin{aligned} var\left(\left(Z-\mu_Z\right)u\right) &= E\left[\left(\left(Z-\mu_Z\right)u\right)^2\right] - \left[E\left(\left(Z-\mu_Z\right)u\right)\right]^2 \\ &= E\left(\left(Z-\mu_Z\right)^2u^2\right) \\ &= E\left(Z-\mu_Z\right)^2 \cdot Eu^2 \\ &= var\left(Z-\mu_Z\right)var(u) \end{aligned}$$

- $var\left(\widehat{\beta}_1^{TSLS}\right)$ increases in $var\left(Z-\mu_Z\right)$ and $var(u)$, and decreases in sample size n and $cov\left(Z,X\right)^2$

General Instrument variable regression

k endogenous regressors, m instrument variables, and r exogenous regressors

The General Instrumental Variables Regression Model and Terminology

Key Concept 12.1

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad (12.12)$$

$i = 1, \ldots, n$, where

- Y_i is the dependent variable;
- u_i is the error term, which represents measurement error and/or omitted factors;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown regression coefficients; and
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$); they are underidentified if $m < k$; and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

Relevance and validity of instruments

- The regression model is
  - **exactly identified** when $m = k$
  - **over-identified** when $m > k$. Allows instrument validity to be tested
  - **under-identified** when $m < k$. The regression model cannot be estimated.

KEY CONCEPT 12.3

THE TWO CONDITIONS FOR VALID INSTRUMENTS

A set of m instruments $Z_{1i}, \ldots, Z_{mi}$ must satisfy the following two conditions to be valid:

1. Instrument Relevance

- *In general*, let $\hat{X}_{1i}^{*}$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and the included exogenous regressors (W's), and let "1" denote the constant regressor that takes on the value 1 for all observations. Then $(\hat{X}_{1i}^{*}, \ldots, \hat{X}_{ki}^{*}, W_{1i}, \ldots, W_{ri}, 1)$ are not perfectly multicollinear.
- *If there is only one* X, then for the previous condition to hold, at least one Z must enter the population regression of X on the Z's and the W's.

2. Instrument Exogeneity

The instruments are uncorrelated with the error term, that is, $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$.

General TSLS

1. Regress each endogenous regressor on all instruments and exogenous regressors:

$$X_{ji} = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_{ji}$$

2. Regress the response variable on predicted values from stage 1 (replacing the endogenous regressors), and the exogenous variables.

$$Y_i = \gamma_0 + \gamma_1 \widehat{X}_{1i} + \cdots + \gamma_k \widehat{X}_{ki} + \gamma_{k+1} W_{1i} + \cdots + \gamma_{k+r} W_{ri} + u_i$$

3. Bias-correct the estimators $\widehat{\gamma}_j$ to obtain (large sample) unbiased estimators $\widehat{\beta}_j^{TSLS}$
4. STATA does all this for you

STATA IV regression,

Demand curve for cigarettes, price instrumented by sales tax (rtaxo) and also cigarette tax (rtaxs)

```
. ivreg lpackpc (lravgprs = rtaxso) lperinc, r

Instrumental variables (2SLS) regression          Number of obs =      96
                                                  F(  2,    93) =   38.16
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.5478
                                                  Root MSE      =   .1656

------------------------------------------------------------------------------
             |               Robust
    lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
   lravgprs |  -1.214456   .2016349    -6.02   0.000    -1.614862   -.8140486
    lperinc |    .248306   .1807971     1.37   0.173    -.1107213    .6073333
      _cons |   9.690355   .6133863    15.80   0.000     8.472292    10.90842
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso

. ivreg lpackpc (lravgprs = rtaxso rtaxs) lperinc, r

Instrumental variables (2SLS) regression          Number of obs =      96
                                                  F(  2,    93) =   49.16
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.5486
                                                  Root MSE      =  .16544

------------------------------------------------------------------------------
             |               Robust
    lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
   lravgprs |  -1.229101   .1545899    -7.95   0.000    -1.536086   -.9221164
    lperinc |   .2568496   .1526548     1.68   0.096    -.0462926    .5599918
      _cons |   9.736457     .51408    18.94   0.000     8.715596    10.75732
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtaxs
------------------------------------------------------------------------------
```

Figure: The results in the book (equations 12.15 and 12.16) are different!

Invalid instruments produce meaningless results.

- Instrument relevance is checked by the regression of exogenous regressors on instruments: is the regression improved when including the instruments together with the exogenous regressors?
  - An instrument Z for X is weak when $cov(X, Z) \approx 0$. Then $cov(Y, Z) = cov(\beta_0 + \beta_1 X + u, Z) = \beta_1 cov(X, Z) \approx 0$ and

$$\beta_1 = \frac{cov(Y, Z)}{cov(X, Z)} \approx \frac{0}{0}$$

  - Weak instruments give weak identification!
  - Weak instruments will also make bad estimates: biased, and highly variable
- Instrument exogeneity testing when $m > k$
  1. $\widehat{u}_i = Y_i - \widehat{Y}_i^{TSLS}$, regress $\widehat{u}_i$ (1) on $Z_1, \cdots, Z_m, W_1, \cdots, W_r$ and (2) on $W_1, \cdots, W_r$
  2. Test wether (1) improves the fit over (2) by an F–test - compare $J = mF$ to the χ^2_{m-k} distribution.

Example

Draft lottery date

1. Angrist (1990) wanted to estimate the effect on wage of an extra year of schooling
   - Schooling and wage are both influenced by talent - an omitted variable which is impossible to measure
   - During the USA-war in Vietnam, students in college could postpone their military service relative to the date drawn in the draft lottery.
     - Students with draft date early in the school year had an extra incentive to take an extra year of shooling.
     - draft date is exogenous (a lottery) and relevant (correlated with years at school) - a successful instrument
   - Joshua D. Angrist: Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records The American Economic Review, Vol. 80, No. 3 (Jun., 1990), pp. 313-336

Early mortality for Europeans in colonies

1. Acemoglu et al wanted to measure the effect of institution quality on economic performance in former colonial countries
   - Good economy allows good institutions, and good institution helps the economy. Causality flows both ways!
   - In colonial time, better institutions were set up in colonies where Europeans wanted to settle, and they wanted to settle if soldiers, missionaries and sailors did not die too rapidly in the early days.
     - Early mortality for soldiers, missionaries and sailors is exogenous to later institutions and economic conditions, and is correlated with settlement and thus institution quality. Good and valid instrument!
   - Acemoglu, D., S. Johnson and J.A. Robinson (2001) The Colonial Origins of Comparative Development: an Empirical Investigation, American Economic Review, Vol. 91, No. 5.

Summary

- $E(Y|\text{ do } X = x)$ is often what we want, but is usually different from $E(Y|\ X = x)$
- Omitted variables, measurement errors in regressors and the regressor X being causally influenced by Y are source of difficulties.
- Instrument variable regression can provide valid inference for $E(Y|\text{ do } X = x)$ etc.
- Instrumentation is more of an art than a science - imagination and critical judgement are needed!
- Reporting results from IV-regression requires good arguments. The reader must be convinced that the instruments are valid: exogenous and relevant.
- Weak instruments are dangerous. It does not help with much data or many weak instruments...