JBS Advanced Quantitative Research Methods Module MPO-1A

Lent 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 3
Endogeneity, IV and SEM

Contents

Instrumental variables: estimators and tests

Summary:

1. OLS estimators are inconsistent if regressors are not exogenous (that is, if x_i is correlated with ε_i). This may happen in presence of *omitted variables*, when these are also correlated with other explanatories. Other possible cases are *measurement errors* and *simultaneous determination* between regressand and regressor.
2. Consequence: it is impossible to distinguish the impact of x on y. If the reason for endogeneity is an omitted variable, the parameter on x will collect the impact of this variable and of the omitted one as well.
3. If instrumental variables are exogenous, however, then IV estimators are consistent (even under heteroscedasticity or autocorrelation).
4. IV estimators are relatively more efficient the more highly correlated the instruments Z are with the explanatory variables. In practice, however, to qualify for exogeneity, instrumental variables have to be relatively weakly correlated with the explanatory variables. Such weak instruments lead to *relatively large variances* on the IV estimator.
5. Practical Procedures: (the steps that go into the canned commands in, for example, R):
   a) Try to use intuition to interpret whether endogeneity might play a role. Possible reasons: *simultaneous correlations* as in demand and supply systems, *omitted variables*, or *measurement errors*.

   b) If it does, find the required number of instruments (new variables) that are exogenous (not correlated with the error: *exclusion restrictions*) and that are likely to be correlated with *the possibly endogenous regressors* (and thus carry information on their variation).

   c) Check whether the proposed instruments are good instruments, that is: sufficiently correlated with the regressors. The rank condition. If endogeneity (of regressors) is weak then OLS may still be considered a suitable method: the bias due to endogeneity may be compensated by the higher efficiency of OLS, compared to the IV.

   d) Check that the number of excluded exogenous regressors (instruments) is equal to or higher than the number of endogenous explanatory variables. Order condition.

   e) If regressors are suspected to be endogenous and instruments are valid then apply IV/2SLS method and compare results with OLS. Observe differences in coefficients and standard errors.

   f) Check whether the proposed instruments are indeed exogenous (Sargan test of the validity of the IVs).

   g) Investigate the suspected endogeneity of regressors. For this purpose: apply Hausman or Durbin-Wu and Hausman tests for the exogeneity of the regressors.
6. Note on point b) above: Looking for instruments. All variables considered exogenous should be considered in the set of instruments (z_j). The constant, should be one of these. Choose as many instruments as non-exogenous (i.e. endogenous) regressors.
7. Note on point e) above: **2SLS**. Two-stage least squares method of estimation.

Step 1. Regress each variable x_j suspected to be endogenous on the set of IV and calculate the predicted variable: $\hat{x}_j$.

Step 2. Regress y on predicted $\hat{x}_j$ and all other exogenous regressors.

8. Note on point f) above: we may apply **Sargan test**.

Step 1. Carry out IV. Calculate residuals: e_{IV}. (These are 'reliable' estimates of the error terms).

Step 2. Perform auxiliary regression $e_{IV} = \sum_j z_j \cdot \gamma_j + \eta$. Retain the value R^2 calculated for this regression. Note that z_j includes all instruments (the constant included).

Step 3. Calculate LM = nR^2 ~ χ^2_{m-k}. Under the null of all instruments being exogenous LM ~ χ^2_{m-k}. (m = no. of instruments, that is the number of variables as z_j; k = no. of regressors). (R^2 corresponds to the auxiliary regression run in Step 2). We're basically testing the null of no instrumental variable being correlated to the errors.

9. Note on point g) above: apply: **Durbin-Wu-Hausman test**. Suppose we have k regressors. The first k-k_o are exogenous, no doubt. The last k_o may be endogenous.

Step 1: Perform a preliminary OLS regression of y on all x. Predict e = residuals.

Step 2: Regress every possible endogenous regressor x_j on the vector of IV (z_j). Predict v_j = estimated residuals corresponding to each of these regressions $\left(v_j = x_j - \sum z_j \cdot \hat{\gamma}_j\right)$. Note that a new set of residuals should be computed for each regressor which is potentially endogenous.

Step 3: Regress e on all regressors (X) (both exogenous ones and potentially endogenous ones), and on all residuals v_j (all of them, calculated with each potentially endogenous variable).

$$e_i = \sum_{j=1}^{j=k} \beta_j \cdot x_{ji} + \sum_{j=k-ko+1}^{j=k} \alpha_j \cdot \hat{v}_{ji} + \eta_i$$

Step 4: Calculate the Lagrange Multiplier statistic: LM = nR^2 ~ χ^2_{k-ko}, under the null of all regressors being exogenous. (k = no. of regressors, ko=no. of regressors potentially endogenous). If we reject the null, then there is endogeneity.

All of the above are canned in R.

Exercise 1. IV regression, Sargan and Hausman test (1)

Use bonds.csv. We will model monthly changes in the US AAA bond rate (y=daaa), in terms of changes in the short-term interest rate (the three-month US Treasury Bill rate=dus3mt), x. $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$. Follow the procedure suggested in point 5.

Answer:

a) General financial (and unobserved)[1] conditions are expected to affect the bond-rate and the short-term interest rate simultaneously. In this case, x_i is not exogenous and OLS is not consistent.

b) In efficient markets all past information is processed in current prices. If this is so, ε_t is uncorrelated with past values of y and x. corr(ε_t , y_{t-j})= corr(ε_t , x_{t-j})= 0; j>0. But time series variables are generally autocorrelated, so x_t and x_{t-1} are correlated. Then, lagged values for x are used as instruments for the x variable. In particular, we'll consider x_{t-1} and x_{t-2}.

c) We want instruments for dus3m, suspected not exogenous. IVs (dus3m, lagged one and two periods) are correlated with dus3m. F-test for the model rejects the null of not being so and each variable is significant.

d) Number of IVs = m = 3 (the two variables plus the constant); number of regressors = k = 2. Then this is correct, as it has to be: m ∃ k.

e) There is an upward bias in the constant and a critical downward bias in the slope when using OLS. IV estimators are, however, less efficient (bigger std errors).

WARNING: If we proceed with the two stages ourselves, we have to be careful to get the standard errors correct in the second step. See the following paragraph.

> A common intuition for the treatment for the endogenous regressor (x_{end}) is as follows (error-in-variables example). As always, the model is (1) $y_i = \alpha + \beta \cdot x_i + u_i$. Imagine $x_{end,i} = x_i + \upsilon_i$, i.e. the variable is observed with error. If we run the model using $x_{end,i}$, the model can be written as: $(2)\, y_i = \alpha + \beta \cdot x_{end,i} + (u_i - \beta \cdot \upsilon_i)$. But note than $x_{end,i} = x_i + \upsilon_i$ is correlated with the error term $(u_i - \beta \cdot \upsilon_i)$, this is why we have to use IV estimation.
>
> Now, the 2SLS method does the following: in the second step we substitute $x_{end,i}$ by a predicted value ($\hat{x}_{end}$), therefore, instead of (1) we estimate (2). The estimator for β in (2) is consistent, but the estimation of the error, $(u_i - \beta \cdot \upsilon_i)$, is biased from the point of view of those in model (1) (u_i), the ones we are interested in – as in (2) they include $\beta \cdot \upsilon_i$. Estimations with R's ivreg overcomes this problem, so the recommendation is: after you understand what it is basically being done, use R's built in commands.

To conclude: use ivreg. It's easier and you get the right std errors!

f) We don't reject the null of the instruments being exogenous. Perform the Sargan test as described in the note on point f) above.

g) Durbin-Wu-Hausman and Hausman tests are performed as described in the note on point g) above. We conclude that the variable is endogenous.

Exercise 2. IV regression, Sargan and Hausman test (2)

Use mroz.csv (from Mroz (1987)), to estimate a wage equation for females.

[1] Note: 'unobserved' does not mean that no-one observes the variable – only that it is not observed by the analyst, and that this component is, therefore, a source of unobserved heterogeneity.

a. Observe the summary statistics looking for relevant information regarding the wage equation in female. Graph a histogram for wage and lwage, and comment on the shape of the distribution.

b. For the remainder of this exercise, retain only the women that work in the data set.

c. Estimate the model: $lwage_i = \beta_0 + \beta_1 \cdot educ_i + \beta_2 \cdot exper_i + \beta_2 \cdot exper_i^2 + u_i$. How would you interpret the coefficients? Do the estimated values make sense? Which kind of problems may these present?

d. Estimate the model by manual 2SLS (do not use ivreg2 yet), using *fatheduc* and *motheduc* as instruments for *educ.* How do the results differ from those obtained with *educ*?

e. Do *fatheduc* and *motheduc* explain *educ*? Are these good instruments?

f. A problem with 2SLS is that the standard errors in the second stage are wrong. The correct standard errors are provided by the **ivreg** command in R. Comment on the differences.

g. Perform the test for overidentifying restrictions, i.e.: whether the instruments are uncorrelated with *u*.

h. Test for endogeneity of *educ*.

Exercise 3. [Homework]

Use crime.csv. In this exercise we consider simulated data (from Heij, et al (2004)) on the relation between police (x) and crime (y), some of the data refer to election years (*election*=1), the other data to non-election years (*election*=0). We want to estimate the effect of police on crime –that is, the parameter β in the model: $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$.

a) Observe the data. Regress *crime* on a constant and *police*. Give a possible explanation of the estimated positive effect.

b) Think of the motivation why the election dummy might serve as an instrument.

c) It can be proved that the IV estimator of β is given by $\left(\frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} \right)$, where $\bar{x}_1$ denotes the sample mean of x over election years and $\bar{x}_0$ over non-election years, and where $\bar{y}_1$ and $\bar{y}_0$ are defined in a similar way. Give an intuitive motivation for this estimator of β.

d) Use the data to estimate β by instrumental variables, using *z* (and a constant) as instruments. Check that the result of c) holds true. Provide an interpretation of the resulting estimate.

e) Perform the Hausman test on the exogeneity of the variable *x*.

Answer:

a) Police seems to have a positive impact on crime both according to the scatter diagram and to OLS estimation. This weird result may be biased by the endogeneity between crime and police, as crime may explain changes in police.

b) The dummy variable *election* is not endogenous, as crime has no impact on election dates. On the other hand, *election* is expected to be correlated with *police*, as a result of the political cycle.

c) In election year the police squads are on average enlarged by the amount $\overline{x}_1 - \overline{x}_0$, in turn, the amount of crime increase by $\overline{y}_1 - \overline{y}_0$. So β_{IV} collects the average increase (or decrease if negative) in crime per increase in the average amount of police.

d) See script-file.

e) The null of exogeneity is rejected at 5%-level.

Poor choices of instruments

In the case of endogeneity occurring because of an omitted variable, is a proxy for the omitted variable a good IV? No it's not, as it'll be correlated with the error. Therefore: do not just swap the instrument with the endogenous variable.

Poor instrumental variables: when z and x are weakly correlated.

If z and u are possibly correlated: $p\lim \hat{\beta}_{IV} = \beta + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$, observe that even if the regressors have not a high correlation with the perturbation, a very low correlation between instruments and regressors (weak instruments) will produce very much biased estimators. This could occur in such a way that even a biased OLS estimation could be preferred as the bias could be lower than the IV's one: $p\lim \hat{\beta}_{OLS} = \beta + Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x}$.

Exercise 4. [Homework]

Use brwght.csv. Model the weight at birth log(bwght) and explain it with the number of cigars smoked by the mother per day (packs). **a)** Estimate using OLS. **b)** Why could this estimation go wrong? **c)** Would the price of cigarettes (cigprice) be a good instrument? **d)** Estimate by IV using cigprice as an instrument. Explain the results.

a) See Script-file.

b) *packs* may be correlated with other unobserved factors affecting health, (for instance, income, if people less wealthy are heavier smokers) originating a problem of endogeneity.

c) Cigarette prices are supposedly not correlated with the weight at birth, but it is correlated (again supposedly) with the consumption of cigars. Therefore it might be a good instrument.

d) When we instrument, the sign of the estimation is positive and the standard deviation huge. The reason is that it is a very weak instrument, as we can see if we regress **packs** on cigprice (In this case the main variable is non-significant). As a result, OLS results look more sensible than IV.

Simultaneous equation models (SEM)

Exercise 5. [HOMEWORK]

Simulating data on macroeconomic consumption (C), disposable income (D), and government expenditures (G). (Heij et al, 2004). The model is

$$\textit{consumption equation}: \quad C_t = \alpha + \beta \cdot D_t + \varepsilon_{1t}$$

$$\textit{income equation}: \quad D_t = C_t + G_t + \varepsilon_{2t}$$

$$\varepsilon_{1t} \sim N(0,\sigma_1); \quad \varepsilon_{2t} \sim N(0,\sigma_2)$$

Here the government expenditures are assumed to be exogenous (that is independent of ε_{1t} and ε_{2t}. Both errors are independent. The parameter of interest is the multiplier (average effect of government expenditures on income).

a) Find the reduced form and define the parameter to be estimated:

b) It can be proved that by estimating β by OLS, with the consumption equation $p\lim(\hat{\beta}_{OLS}) = \beta + \frac{(1-\beta)\cdot\sigma_1^2}{\sigma_G^2 + \sigma_1^2 + \sigma_2^2}$. What can you say about the bias of estimating this parameter by OLS? In which variables does it depend on?

c) Simulate n=100 observations from this model with α=0, β=0.5 (therefore the multiplier is equal to 2), the perturbations are distributed as standard normals, G is distributed normal with mean 10 and variance 1. Applying the formula above: plim(b) = 0.5 + 0.5/3 = 0.67, and the multiplier would be around 3. Note that estimating the consumption equation by OLS would overestimate the real multiplier, which is 2. (Try again with a sample of 50 and 10000 and note the differences with the predicted values).

d) Apply instrumental variables to obtain better estimators. Report the value for the multiplier estimated.

e) In SEM, IV are applied to each equation with the exogenous variables in the system as IVs. This is called the 2SLS (two-stage least squares method). Then the order condition must be satisfied: the number of IVs that do not appear in the equation should be at least as large as the number of endogenous regressors in the equation. k – k_i >= m_i ; where k = number of exogenous variables; k_i = number of exogenous variables that appear in the *i*th equation, m_i = number of endogenous variables that appear in the *i*th equation. Based on this property, explain the exclusion restrictions.

Answer:

a) $\textit{reduced form}: \quad D_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta}\cdot G_t + \frac{\varepsilon_{1t}+\varepsilon_{2t}}{1-\beta}$

The multiplier is then: $\frac{1}{1-\beta}$

b) The inconsistency of OLS is relatively small if σ_2^2 or σ_G^2 are large compared to σ_1^2. That is if the variation in the error in the consumption equation is small compared to the variation of public expenditure and of variations in the error in the income equation.

e) k – k_i >= m_i, following from : (k – k_i) + (m-m_i - 1) >= m-1

(k – k_i) = number of exogenous variables that do not appear in the *i*th equation (m-m$_i$ -1) = number of endogenous variables that do not appear in the *i*th equationboth added = number of excluded variables from the ith equation. It should be at least equal to the number of endogenous variables of the model (m) minus 1.

In the case of a two-equation system, identification of one equation is quite simple. Take, for example, the first equation. For it to be identified, there must be at least one exogenous variable excluded from the first equation that appears with a nonzero coefficient in the second equation.

Exercise 6.

Use oranges.csv. Data correspond to 50 years in the US market for oranges (1910-59). The data are taken from Neerlove and Waugh (1961). The variables are the quantity traded (Q), the price (P), real disposable income (RI), current advertisement expenditures (AC), and past advertisement expenditures (AP, averaged over the past ten years). First, we assume that the supply Q is fixed and that the price is determined by demand via the price equation $\log(P_t) = \alpha + \gamma \cdot \log(Q_t) + \beta \cdot \log(RI_t) + \varepsilon_t$

a. Estimate the price equation using OLS. Test the null hypothesis that price elasticity $(\gamma = -1)$.

b. Estimate the price equation also by IV, using log(AC$_t$) and log(AP$_t$) as instruments for log(Q). Test again the null hypothesis of unit price elasticity.

c. Perform the Hausman test for the exogeneity of log(Q$_t$) in the price equation.

d. Investigate the quality of the instruments – that is, whether they are sufficiently correlated with log(Q$_t$) and uncorreated with the price shocks ε_t (take the IV residuals as estimates of the shocks).

e. Answer questions b,c and d also for the n=45 observations obtained by excluding the data over the period 1942-6.

Next we consider the simultaneous model for price and quantity described by the following two equations. We exclude the two advertisement variables (AC and AP) in f and g.

$$(demand) \quad \log(P_t) = \alpha + \gamma \cdot \log(Q_t) + \beta \cdot \log(RI_t) + \varepsilon_{1t}$$

$$(supply) \quad \log(P_t) = \alpha + \gamma \cdot \log(Q_t) + \varepsilon_{2t}$$

f. Is the demand equation identified? Estimate this equation by OLS, and motivate your choice.

g. Is the supply equation identified? Estimate this equation by a method that you find most appropriate, and motivate your choice.

Exercise 7. [HOMEWORK]

Use smoke.csv.

a) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness, or productivity effect) is $\log(income) = \beta_0 + \beta_1 \cdot cigs + \beta_2 \cdot educ + \beta_3 \cdot age + \beta_4 \cdot age^2 + u_1$, where *cigs* is number of cigarettes smoked per day, on average. How would you interpret β_1?

b) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is:

$$cigs = \gamma_0 + \gamma_1 \cdot \log(income) + \gamma_2 \cdot educ + \gamma_3 \cdot age + \gamma_4 \cdot age^2 + \gamma_5 \cdot \log(cigprc) + \gamma_6 \cdot restaurn + u_2$$

where cigpric is the price of a pack of cigarettes (in cents), and restaurn is a binary variable equal to unity if the person lives in a state with restaurant smoking restrictions. Assuming these are exogenous to the individual, what signs would you expect for γ_5 and γ_6.

c) Under what assumption is the income equation from part a identified?

d) Estimate the income equation by OLS and discuss the estimate of β_1.

e) Estimate the reduced form for cigs. (Recall that this entails regressing cigs on all exogenous variables.) Are log(cigpric) and restaurn significant in the reduced form?

f) Now, estimate the income equation by 2SLS. Discuss how the estimate of β_1 compares with the OLS estimates.

g) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?

Answer:

a) Assuming the structural equation represents a causal relationship, $100.\beta_1$ is the approximate percentage change in income if a person smokes one more cigarette per day.

b) Since consumption and price are, ceteris paribus, negatively related, we expect $\gamma_5 < 0$. Similarly, everything else equal, restaurant smoking restrictions should reduce cigarette smoking, so $\gamma_6 < 0$.

c) We have a two-equation system. In this case, for the first equation to be identified there must be at least one exogenous variable excluded from the first equation that appears with a nonzero coefficient in the second equation. In our example, we need γ_5 or γ_6 to be different from zero. That is, we need at least one exogenous variable in the *cigs* equation that is not also in the log(*income*) equation.

d) The coefficient on *cigs* implies that cigarette smoking causes income to increase (!), although the coefficient is not statistically different from zero. Remember, OLS ignores potential simultaneity between income and cigarette smoking, and therefore we should assume that the estimation is biased.

e) When the model is expressed in 'reduced form' each endogenous variable (in this part, log(income)) is regressed on all exogenous variables and none of the endogenous ones.

While log(*cigpric*) is very insignificant, *restaurn* has the expected negative sign and a *t* statistic of about –2.47. People living in states with restaurant smoking restrictions smoke almost three fewer cigarettes, on average, given education and age. (remember the *ceteris paribus* assumption). We could drop log(*cigpric*) from the analysis (it is clearly not significant) but we leave it in. The *F* test for joint significance of log(*cigpric*) and *restaurn* yields *p*-value .044.

f) Now the coefficient on *cigs* is negative (as expected) and almost significant at the 10% level against a two-sided alternative. The estimated effect is very large: each additional cigarette someone smokes lowers predicted income by about 4.2%. Of course, as we are working with IVs, the 95% CI for β_{cigs} is very wide.

g) State level cigarette prices and restaurant smoking restrictions could be considered endogenous in the income equation. Incomes are known to vary by region, as do restaurant smoking restrictions. It could be that in states where income is lower (after controlling for education and age), restaurant smoking restrictions are less likely to be in place.

Exercise 8. [HOMEWORK]

Use openness.csv to check whether more open countries should have lower inflation rates (Romer, 1993). Openness is assessed as the average share of imports in gross domestic product. The system he has in mind is:

(1) $\inf_t = \beta_{10} + \alpha_1 \cdot open + \beta_{11} \cdot \log(pcinc_t) + u_{t1}$

(2) $open_t = \beta_{20} + \alpha_2 \cdot \inf + \beta_{21} \cdot \log(pcinc_t) + \beta_{22} \cdot \log(land_t) + u_{t1}$

where pcinc is 1980 per capita income in US dollars (assumed exogenous), and land is the land area of the country, in square miles, and also assumed exogenous. Equation 1 is the one of interest, with the hypothesis that $\alpha_1 < 0$. (More open economies have lower inflation rates). The second equation reflects the fact that the degree of openness might depend on the average inflation rate as well as other factors. Among these log(land), as ceteris paribus, smaller countries are likely to be more open ($\beta_{22} < 0$).

a) To confirm the last assertion, estimate the reduced form for open. Is the first equation identified? Estimate it using a constant and log(land) as IVs (2SLS).

b) Because log(pcinc) is insignificant in both estimations so far [CHECK THIS], drop it from the analysis. Estimate by OLS and IV without log(pcinc). Do any important conclusions change?

c) Still leaving log(pcinc) out of the analysis, is land or log(land) a better instrument for open? (Hint: regress open on each of these separately and jointly).

d) Add the dummy variable oil (indicative of the country being an oil-producer) to the original equation in a and treat it as exogenous. Estimate the equation by IV. Does being an oil producer have a ceteris paribus effect on inflation?

Answer:

a) Land has an important negative impact on openness, as Romer (1993) asserts. The first equation is identified, if and only if $\beta_{22} \neq 0$. Equation 2 is not identified but we're interested in equation 1.

b) The IV estimate with log(*pcinc*) in the equation is .338, which is very close to .333. Therefore, dropping log(*pcinc*) makes little difference.

c) Subject to the requirement that an IV is exogenous, we want an IV that is as highly correlated as possible with the endogenous explanatory variable. If we regress *open* on *land* we obtain R^2 = .095. The simple regression of *open* on log(*land*) gives R^2 = .448. Therefore, log(*land*) is much more highly correlated with *open*. Further, if we regress open on log(*land*) and *land* we get that while log(*land*) is very significant, *land* is not, so we might as well use only log(*land*) as the IV for open.

d) Being an oil producer is estimated to reduce average annual inflation by over 6.5 percentage points, but the effect is not statistically significant. This is not too surprising, as there are only seven oil producers in the sample.

Exercise 9.

Use cement.csv.

a) A static (inverse) supply function for the monthly growth in cement price (gprc) as a function of growth in quantity (gcem) is $gprc_t = \alpha_1 \cdot gcem_t + \beta_0 + \beta_1 \cdot gprcpet_t + \beta_2 \cdot feb_t + ... + \beta_4 \cdot dec_t + u_t^s$, where gprcpet (growth in the price of petroleum) is assumed to be exogenous and feb, … , dec are monthly dummy variables. What signs do you expect for α_1 and β_1? Estimate the equation by OLS. Does the supply function slope upward?

b) The variable gdefs is the monthly growth in real defense spending in the United States. What do you need to assume about gdefs for it to be a good IV for gcem? Test whether gcem is partially correlated with gdefs. (Do not worry about possible serial correlation in the reduced form.) Can you use gdefs as an IV in estimating the supply function?

c) Shea (1993) argues that the growth in output of residential (gres) and nonresidential (gnon) construction are valid instruments for gcem. The idea is that these are demand shifters, that should be roughly uncorrelated with the supply error u_t^s. Test whether gcem is partially correlated with gres and gnon; again, do not worry about serial correlation in the reduced form.

d) Estimate the supply function, using gres and gnon as IVs for gcem What do you conclude about the static supply function for cement? [The dynamic supply function is, apparently, upward sloping; see Shea (1993)].

Answer:

a) For α_1 we should expect a positive sign as it is a supply function, β_1 should also be positive indicating the positive effect on cement prices of an increase in one of the main inputs. If we estimate the inverse supply function by OLS several of the monthly dummy variables are statistically very significant, but their coefficients are not of direct interest here. The estimated supply curve slopes *down*, not up, and the coefficient on $gcem_t$ is very statistically significant (*t*- statistic = 4.87).

b) For gdefs to be a good IV for gcem we need to assume first, that gcem is excluded from the supply equation, which seems clear, as there shouldn't be any contemporaneous effect from the expenditure in defence in the supply of cement. So we could think that as an instrumental variable, it is exogenous. To be a good IV, still we need $gdefs_t$ to have a nonzero coefficient in the reduced form for $gcem_t$.

$gcem_t = \beta_0 + \beta_1 gdefs_t + \beta_2 gprcpet_t + \beta_3 feb_t + \; + \beta_{13} dec_t + v_t,$

Then identification requires $\beta_1 \neq 0$. For this we do not have to assume anything, though. Simply run OLS.

When we run this regression, β_1 = - 1.054 with a *t* statistic of about –0.294. Therefore, we cannot reject H_0: $\beta_1 = 0$ at any reasonable significance level, and we conclude that $gdefs_t$ is not a useful IV for $gcem_t$ (even if $gdefs_t$ is exogenous in the supply equation).

c) Now the reduced form for *gcem* is
$gcem_t = \beta_0 + \beta_1 gres_t + \beta_2 gnon_t + \beta_3 gprcpet_t + \beta_4 feb_t + \ + \beta_{14} dec_t + v_t,$
and we need at least one of β_1 and β_2 to be different from zero. In fact, $\hat{\beta}_1 = .136$, $t(\hat{\beta}_1) = .984$ and $\hat{\beta}_2 = 1.15$, $t(\hat{\beta}_2) = 5.47$. So $gnon_t$ is very significant in the reduced form for $gcem_t$, and we can proceed with IV estimation.

d) We use both $gres_t$ and $gnon_t$ as IVs for $gcem_t$ and apply 2SLS, even though the former is not significant in the reduced form. In the estimated supply function the coefficient on $gcem_t$ is still negative. However, it is only about one-fourth the size of the OLS coefficient, and it is now very insignificant. At this point we would conclude that the static supply function is horizontal (with *gprc* on the vertical axis, as usual). Shea (1993) adds many variables, lags of $gcem_t$ and others, obtaining a positive long run slope.

Exercise 10.

Use the data set in fish.csv, which comes from Graddy (1995), to estimate a demand function for fish.

a) Assume that the demand equation can be written in equilibrium for each time period as: $\log(totqty_t) = \alpha_1 \cdot \log(avgprc_t) + \beta_{10} + \beta_{11} \cdot mon_t + ... + \beta_{14} \cdot thurs_t + u_{t1}$, so that demand is allowed to differ across days of the week. Treating the price variable as endogenous, what additional information do we need to consistently estimate the demand-equation parameters?

b) The variables wave2 and wave3 are measures of ocean wave heights over the past several days. What two assumptions do we need to make in order to use wave 2 and wave 3 as IV for log(avgprc) in estimating the demand equation?

c) Regress log(avgprc) on the day-of-the-week dummies and the two wave measures. Are wave2 and wave3 jointly significant? What is the p-value of the test?

d) Now, estimate the demand equation by 2SLS. What is the 95% confidence interval for the price elasticity of demand? Is the estimated elasticity reasonable?

e) Given the supply equation evidently depends on the wave variables, what two assumptions would we need to make in order to estimate the price elasticity of supply?

f) In the reduced form equation for log(avgprc) are the day-of-the-week dummies jointly significant? What do you conclude about being able to estimate the supply elasticity?

Answer:

a) To estimate the demand equations, we need at least one exogenous variable that appears in the supply equation.

b) For $wave2_t$ and $wave3_t$ to be valid IVs for $\log(avgprc_t)$, we need two assumptions. The first is that these can be properly excluded from the demand equation. This is arguable, as wave heights are determined partly by weather, and demand at a local fish market could depend on weather. The second assumption is that at least one of $wave2_t$

and $wave3_t$ appears in the supply equation. There is indirect evidence of this in part c), as the two variables are jointly significant in the reduced form for log($avgprc_t$).

c) The variables $wave2_t$ and $wave3_t$ are jointly very significant: F = 19.1, p-value = zero to four decimal places.

d) The 95% confidence interval for the demand elasticity is roughly -1.46 to -0.17. The point estimate, -0.82, seems reasonable: a 10 percent increase in price reduces quantity demanded by about 8.2%.

e) To estimate the supply elasticity, we would have to assume that the day-of-the-week dummies do not appear in the supply equation, but they do appear in the demand equation. Part (c) provides evidence that there are day-of-the-week effects in the demand function.

f) Unfortunately, in the estimation of the reduced form for log($avgprc_t$) in part (c), the variables *mon*, *tues*, *wed*, and *thurs* are jointly insignificant [F(4,90) = .53, p-value = .71.] This means that, while some of these dummies seem to show up in the demand equation, things cancel out in a way that they do not affect equilibrium price, once *wave2* and *wave3* are in the equation. So, without more information, we have no hope of estimating the supply equation.

For more exercises check: http://www.ats.ucla.edu/stat/examples/greene/

References

- Graddy (1995), 'Testing for Imperfect Competition at the Fulton Fish Market', *Rand Journal of Ecoomics* 26, 75-92.
- Mroz (1987), 'The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions', *Econometrica* 55, 765-799
- Neerlove and Waugh (1961) 'Advertising without Supply Control: Some Implications of a Study of the Advertising of Oranges', *Journal of Farm Economics*, 43, 813-67.
- Romer, D. (1993), "Openness and Inflation: Theory and Evidence", *Quarterly Journal of Economics, 108, 869-903.*

These notes have benefited from:

- Dougherty (2002), *Introduction to Econometrics.*
- Heij et al. (2004), *Econometric methods with applications in business and economics.*
- Wooldridge (2006), *Introductory Econometrics.*