HW 7 Solution

Chunyu Qu

12/6/2021

I. Summary of Lewbel 2012 - internal instrumental variable

The three main endogeneity resources for linear regression is unobserved variable, measurement error, and simultaneity. With instrumental variable method we can identify the reduced model. However, when IV is not available, usually there is no good way of doing a consistent estimation. Especially, given some heteroscedasticity we can still use this method to identify the parameters. Lewbel's paper demonstrates how higher order moment restrictions can be used to tackle endogeneity in triangular systems. Without going into too much detail (interested readers can consult Lewbel's paper), this method is like the traditional two-stage instrumental variable approach, except the first-stage exclusion restriction is generated by the control, or exogenous, variables which we know are heteroskedastic (interested practitioners can test for this in the usual way, i.e. a White test).

The procedure can be done with the following steps:

- Step 1. Generate the reduced form errors by regressing the endogenous variable Y_i on control variables $x_j, j = 1, 2, ...m$. Obtain the reduced form errors $\hat{v_2}$.

- Step 2. Generate the IVs by creating $\tilde{x_i} = [x_i - \bar{x_i}]\hat{v_2}$, $i = 1, 2, ...n$
- Step 3. Regress Y_i on controls and the generate IVs, for the first stage unrestricted model.
- Step 4. Regress Y_i only on controls, for the restricted model.
- Step 5. Compute the Fstat to determine if the IVs are sufficiently strong.
- Step 6. Regress Y_i on controls and IV errors, for the second stage.

II. PS7 Solution

We have a simultaneous system of three variabels, $Y_1(growth)$ indicate he average growth rate, $Y_2(cored)$ indicates the proportion of households that reporting paying a bribe at educational institutions, and $Y_3(jud)$ is the proportion of households that reporting paying a bribe at legal institutions.

a.Under what conditions can you identify each equation? Describe the approach to estimating each equation using 2SLS.

Answer:

First of all, we have exogenous controls for each equation. The problem is that our simultaneous explanatory variables are endogenous (e.g. $E[Y_2u_1] \neq 0, E[Y_3u_1] \neq 0$ for equation 1, and similar for the other two equations). As long as each equation contains an IV for each of the endogenous variables that satisfies the validity requirement (uncorrelated with u_g) and is sufficiently partially correlated with each endogenous variable, each equation can be identified.

b.Ignoring any potential endogeneity issues above, estimate Equation (1) via OLS.* Interpret how each type of corruption impacts expected average economic growth.

Answer:

```
# Import the data
data = read.csv("corrupt.csv",header = TRUE)
head(data)
```

```
##   Country.Code    Country CPI2010    AvgGrowth AvgPopGrowth GDPPCInit
## 1            9  Argentina     2.9  0.014284966  0.007730928  5.759213
## 2           10    Armenia     2.6 -0.006320365  0.014284966  3.194682
## 3           12 Australia      8.7  0.016546637  0.016546637 11.913727
## 4           13    Austria     7.9  0.003178666  0.003178666  5.759237
## 5           14 Azerbaijan     2.4  0.012217873  0.012217873  4.794606
## 6           19    Belarus     2.5 -0.003358280  0.022178139  6.897408
##        cored     edext    corjud    judext
## 1  0.5376344  7.776670  5.769231 29.112662
## 2 12.2448980 56.300000 16.949153 49.200000
## 3  1.6528926  2.647059 10.144927  6.764706
## 4  3.3333333  2.487562  3.478261  2.487562
## 5 30.0613497 32.023576 43.820225 20.432220
## 6 13.4653465 10.300000 18.556701 10.700000
```

```
# Run the OLS
lm1 = lm(AvgGrowth ~ AvgPopGrowth + GDPPCInit + cored + corjud, data = data)
summary(lm1)
```

```
##
## Call:
## lm(formula = AvgGrowth ~ AvgPopGrowth + GDPPCInit + cored + corjud,
##     data = data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0242373 -0.0046804  0.0008369  0.0061064  0.0159647
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0136615  0.0033405   4.090 0.000142 ***
## AvgPopGrowth  0.1337869  0.0902266   1.483 0.143839
## GDPPCInit    -0.0011017  0.0003099  -3.555 0.000787 ***
## cored        -0.0003339  0.0001720  -1.941 0.057357 .
## corjud        0.0003743  0.0001356   2.761 0.007810 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009116 on 55 degrees of freedom
## Multiple R-squared:  0.4001, Adjusted R-squared:  0.3565
## F-statistic: 9.172 on 4 and 55 DF,  p-value: 9.452e-06
```

Using population growth and gdp per capita as additional controls we produce the above results. For a one unit increase in educational corruption, expected average growth decreases by .03339% holding other

variables in the model fixed. For a one unit increase in judicial corruption, expected average growth increases by .03743% holding other variables in the model fixed.

c.Now assume we observe an potential IV edext for cored. The variable edext measures the percentage of housholds that believe corruption in education is an extreme problem. In your opinion is this IV valid? Explain carefully.

Answer:

This potential IV is valid as long as edext has no direct impact on economic growth or is unrelated to unobserved variables contained in u1, and is correlated with cored. Given we have excluded other variables that likely influence economic growth and are likely to be correlated to perceptions of corruption, this variable is not likely to be a valid IV.

d.estimate Equation (1) via 2SLS using edext as an IV for cored. Is the IV sufficiently strong? How do you know?

Answer:

Use the control function approach to estimate the structural model of choice as it allows for an equivalent 2SLS estimate and an endogeneity test. In the following the first stage test, $tstat = 4.586$, and so $Fstat = 4.586^2 = 21.031 > 10$ so the IV is sufficiently strong.

```
# 1st stage test
lm2 = lm(cored ~ AvgPopGrowth + GDPPCInit + edext, data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = cored ~ AvgPopGrowth + GDPPCInit + edext, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.803  -5.528  -0.612   3.361  49.697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.6920     4.2664   1.334   0.1876
## AvgPopGrowth 134.4688   109.9422   1.223   0.2264
## GDPPCInit     -0.8490     0.3618  -2.347   0.0225 *
## edext          0.5888     0.1284   4.586 2.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 56 degrees of freedom
## Multiple R-squared:  0.4361, Adjusted R-squared:  0.4059
## F-statistic: 14.44 on 3 and 56 DF,  p-value: 4.412e-07
```

Then we use this IV to generate the IV estimates

```
# IV estimate
v2hat = residuals(lm2)
lm3 = lm(AvgGrowth ~ AvgPopGrowth + GDPPCInit + cored + v2hat,data = data)
summary(lm3)
```

```
##
## Call:
## lm(formula = AvgGrowth ~ AvgPopGrowth + GDPPCInit + cored + v2hat,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.021317 -0.006921  0.001518  0.006930  0.018028
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0225771  0.0043176   5.229 2.72e-06 ***
## AvgPopGrowth  0.1598525  0.0942965   1.695   0.0957 .
## GDPPCInit    -0.0017352  0.0003905  -4.443 4.33e-05 ***
## cored        -0.0002569  0.0001827  -1.406   0.1652
## v2hat         0.0004440  0.0002143   2.072   0.0429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009368 on 55 degrees of freedom
## Multiple R-squared:  0.3665, Adjusted R-squared:  0.3204
## F-statistic: 7.953 on 4 and 55 DF,  p-value: 3.921e-05
```

e.Interpret the marginal effect of educational corruption on average economic growth. Is this marginal effect statistically different from zero?

Answer:

Given the results in the IV estimate, a one unit increase in educational corruption, average economic growth decreases by .02569% holding other variables constant. From Listing 3 we can see that the p-value on educational corruption exceeds .05 and so the marginal effect is not statistically different from zero.

f.Now test whether cored is endogenous in Equation (1). What do you conclude?

Answer:

Given part e, we can see the coefficient on $\hat{v_2}$ is statistically different from zero thus educational corruption is endogenous in the structural equation. This conclusion is only valid if the proposed IV is in-fact valid.

g.Suppose you think that the proposed IV above is not a valid IV. How else can you identify Equation (1) under the assumption that cored is still endogenous in the structural equation? Be specific.

Answer:

In the case in which we have no valid IV available for educational corruption we can rely upon heteroskedasticity in the reduced from equation for educational corruption to help identify Equation (1). If we refer to v2 as the population reduced form error for educational corruption then we require the following conditions to obtain identification of Equation (1):

$$E[z_1' u_1] = 0, E[z_1' v_2] = 0$$

$$cov(z_1, u_1 u_2) = 0, cov(z_1, v_2^2) \neq 0$$

where $z_1 = z_2$ as we have no external IVs available to us.

h.Now implement the approach of Lewbel (2010) to estimate the model via 2SLS using generated IVs Are the instruments sufficiently strong?

Answer:

The generated IV is

$$(z - \bar{z})\hat{\epsilon_2}$$

To utilize the approach of Lewbel (2012) we must take a series of steps.

- Step 1. Generate the reduced form errors by regressing cored on our control variables which are AvgPopGrowth and GDPPCInit in our case. Obtain the reduced form errors $\hat{v_2}$.
- step 2. Generate the IVs by creating $\tilde{z_1} = [AvgPopGrowth - \bar{AvgPopGrowth}]\hat{v_2}$, and $\tilde{z_2} = [GDPPCInit - \bar{GDPPCInit}]\hat{v_2}$.
- Step 3. Regress cored on controls and the generate IVs, for the first stage unrestricted model.
- Step 4. Regress cored only on controls, for the restricted model.
- Step 5. Compute the Fstat to determine if the IVs are sufficiently strong.
- Step 6. Regress Y on controls and IV errors, for the second stage.

Using these generated IVs we can obtain the following first-stage regression results:

```
# Generate z1 and z2
# Step 1
lm4 = lm(cored ~ AvgPopGrowth + GDPPCInit, data=data)
r_v2hat = residuals(lm4)

# Step 2
z1 = (data$AvgPopGrowth-mean(data$AvgPopGrowth))*r_v2hat
z2 = (data$GDPPCInit-mean(data$GDPPCInit))*r_v2hat

# Obtain the first stage result
lm5 = lm(cored ~ AvgPopGrowth + GDPPCInit+z1+z2, data=data )
summary(lm5)
```

```
##
## Call:
## lm(formula = cored ~ AvgPopGrowth + GDPPCInit + z1 + z2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8136  -6.9029  -0.3172   6.2060  25.7157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.69578    2.78400   4.919 8.25e-06 ***
## AvgPopGrowth 143.04719  100.98148   1.417  0.16225
## GDPPCInit     -0.78324    0.28216  -2.776  0.00751 **
## z1             5.40200    7.79335   0.693  0.49113
## z2            -0.17479    0.02541  -6.878 5.99e-09 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.947 on 55 degrees of freedom
## Multiple R-squared:  0.6454, Adjusted R-squared:  0.6196
## F-statistic: 25.03 on 4 and 55 DF,  p-value: 7.756e-12
```

What is relevant from this regression is the sum of squared residuals (from the unrestricted model)

$$SSR_{ur} = se^2 \times df = 8.947^2 \times 55 = 4402.684$$

which will help us compute the first-stage F-stat for our generated IVs $\tilde{z_1}, \tilde{z_2}$.

The results for the restricted model (excluding the generated IVs) are presented below:

```
lm6 = lm(cored ~ AvgPopGrowth + GDPPCInit, data=data )
summary(lm6)
```

```
##
## Call:
## lm(formula = cored ~ AvgPopGrowth + GDPPCInit, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.425  -8.571  -3.436   8.232  48.155
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.2857     3.7955   4.818 1.11e-05 ***
## AvgPopGrowth 112.7520   127.6871   0.883 0.380927
## GDPPCInit     -1.4947     0.3874  -3.859 0.000293 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 57 degrees of freedom
## Multiple R-squared:  0.2244, Adjusted R-squared:  0.1971
## F-statistic: 8.244 on 2 and 57 DF,  p-value: 0.0007167
```

Give the results above, we can compute the sum of the squared residuals for the restricted model

$$SSR_r = se^2 \times df = 13^2 \times 57 = 9633$$

Using the formula for the first stage F-stat which is just

$$Fstat = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(9633 - 4402.684)/2}{4402.684/55} = 32.67$$

which suggests that the generated IVs are sufficiently strong. Using these IVs we can generate the control function results which are presented below:

```
v2_hat_2 = residuals(lm5)
lm7 = lm(AvgGrowth ~ AvgPopGrowth+GDPPCInit+cored+v2_hat_2, data=data)
summary(lm7)
```

```
##
## Call:
## lm(formula = AvgGrowth ~ AvgPopGrowth + GDPPCInit + cored + v2_hat_2,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024079 -0.006195  0.001483  0.007950  0.017901
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.614e-02  3.754e-03   4.301 7.03e-05 ***
## AvgPopGrowth  1.202e-01  9.665e-02   1.243  0.21897
## GDPPCInit    -1.209e-03  3.524e-04  -3.431  0.00115 **
## cored         9.491e-05  1.344e-04   0.706  0.48301
## v2_hat_2     -6.355e-05  1.988e-04  -0.320  0.75039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009718 on 55 degrees of freedom
## Multiple R-squared:  0.3183, Adjusted R-squared:  0.2687
## F-statistic: 6.419 on 4 and 55 DF,  p-value: 0.0002592
```

Using the generated IVs the impact of educational corruption on expected growth turns positive however it is not statistically significant. These results are only valid under the identification assumptions required stated in part g.