Asymptotic Properties and simulation in gretl

Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

Outline

Classical Assumptions

Gauss-Markov Assumptions:

- A1: Linearity: $y = \beta + \beta_1 x_1 + ... + \beta_k x_k + v$
- A2: Random Sampling
- A3: Conditional Mean Independence:
  $E[y\,|\mathbf{x}] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
- A4: Invertibility of Variance-covariance Matrix
- A5: Homoskedasticity: $Var[v\,|\mathbf{x}] = \sigma^2$

Normality

- A6: Normality: $y\,|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \sigma^2)$

Asymptotic Properties for OLS (1/2)

Consistency Under Gauss-Markov A1-A4, $plim(\hat{\beta}_j) = \beta_j$

Asymptotic normality (CLT): strong version

- Under Gauss-Markov A.1 to A.5:

$$n^{1/2} \frac{\hat{\beta}_j - \beta_j}{\sigma / a_j} \rightarrow N(0,1) \text{ as } n \rightarrow \infty \text{ where } {a_j}^2 = plim\left(\frac{1}{n}\sum_i \hat{res}_{ji}{}^2\right)$$

Asymptotic efficiency

- Under Gauss-Markov A.1 to A5, OLS is asymptotically efficient in the class of linear estimators

Asymptotic Properties for OLS (2/2)

Asymptotic normality (CLT): weak version

- Under A.1 to A.4:

$$n^{1/2}\left(\hat{\beta}_j - \beta_j\right) \rightarrow N\left(0, n * Avar(\hat{\beta}_j)\right) \qquad \text{as } n \rightarrow \infty$$

- but OLS is not longer asymptotically efficient

- From the CLTs

$$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \rightarrow N(0, 1) \text{ as } n \rightarrow \infty$$

where $plim(se(\hat{\beta}_j)) = \sqrt{Avar(\hat{\beta}_j)}$

Suppose that A3 does not hold

- $y = \beta_0 + \beta_1 x + u$ but $cov(x,u) \neq 0$
- OLS is such that $\hat{cov}_N(x, y - \hat{\beta}_0 - \hat{\beta}_1 x) = 0 \rightarrow \left\{ \hat{\beta}_0, \hat{\beta}_1 \right\}$ is consistent with a false property

Example: $wages = \beta_0 + \beta_1 education + u$

- those with higher ability are more likely to go to college and have higher wages: $cov(educ, u) \neq 0$
- $\hat{\beta}_1$ would overestimate the effect of going to college by the effect of ability on education

- we want to use in the sample a property which is true for the population

Instruments

$$y = \beta_0 + \beta_1 x + u$$

$$cov(x,u) \neq 0$$

- An instrument z is a variable whose influence on the dependent variable is only via a control
  - z is relevant in the sense that it correlates with controls: $cov(x,z) \neq 0$
  - z is exogenous in the sense that controls capture all its effects on the dependent variable: $cov(u,z) = 0$
- each exogenous control is an instrument of itself

IV Estimation: The Basic Idea

$$y = \beta_0 + \beta_1 x + u$$

$$cov(x,u) \neq 0 \text{ (OLS is inconsistent)}$$

$$cov(x,z) \neq 0 \text{ (z is relevant)}$$

$$cov(z,u) = 0 \text{ (z is exogenous)}$$

$$cov(y,z) = \beta_1 cov(x,z) \Rightarrow \beta_1 = \frac{cov(y,z)}{cov(x,z)}$$

we use in the sample a property which is true for the population

$$\hat{\beta}_1^{IV} = \frac{\hat{cov}_N(y_i,z_i)}{\hat{cov}_N(x_i,z_i)}$$

IV Estimation in the General Case

$$y_1 = \beta_0 + \beta_1 z_1 + \beta_2 y_2 + u$$

$$cov(y_2, u) \neq 0$$

- $z_1$is a set of k_1 exogenous variables: $cov(z_1, u) = 0$
- y_2 is a set of k_2 endogenous variables, but there is an instrument for each endogenous variable in y_2, $cov(z_2, u) = 0$
- the system of $k_1 + k_2 + 1$ linear equations

$$\hat{cov}_N\left(z_{1i}, y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 z_{1i} + \hat{\beta}_2 y_{2i}\right) = 0$$
$$\hat{cov}_N\left(z_{2i}, y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 z_{1i} + \hat{\beta}_2 y_{2i}\right) = 0$$
$$\hat{mean}_N\left(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 z_{1i} + \hat{\beta}_2 y_{2i}\right) = 0$$

uniquely identifies $\left\{\hat{\beta}_0^{IV}, \hat{\beta}_1^{IV}, \hat{\beta}_2^{IV}\right\}$

2SLS Assumptions

Gauss-Markov Assumptions

- 2SLS1: Linearity: $y = \beta + \beta_1 x_1 + ... + \beta_k x_k + v$
- 2SLS2: Random Sampling
- 2SLS3: Exogeneity: $cov(u, z) = 0$
- 2SLS4: Rank condition: (i) there are no perfect linear relations among the instruments. (ii) The invertibility (relevance) condition holds.
- 2SLS5: Homoskedasticity: $var[v|\mathbf{z}] = \sigma^2$

2SLS Large Sample Results

Theorem

Under 2SLS1-2SLS4, 2SLS is consistent

Theorem

Under 2SLS1-2SLS5, 2SLS is asymptotically normal and asymptotically efficient in the class of IV estimators

Theorem

Under 2SLS1-2SLS4, 2SLS is consistent and asymptotically normal

Some Properties of 2SLS

- IV standard errors tend to be larger than OLS standard errors
- the stronger the correlation between z and x, the smaller the IV standard errors
- getting non-significant results using IV may simply be a problem of "poor instruments"

Testing after 2SLS

- t-tests: under $H_0 : \beta_j = 0 \Rightarrow t = \frac{\hat{\beta}_j^{IV}}{se(\hat{\beta}_j^{IV})} \xrightarrow{a} N(0,1)$
- it is possible to test for multiple linear hypothesis
- Hausman test for endogeneity H_0: OLS is consistent
- a t-test for endogeneity:
  - First step: regress y_2 on all z_1 and z_2 and compute residual $\hat{v}$
  - Second step: OLS y_1 on z_1 and y_2 AND $\hat{v}$. Under the null, the slope for $\hat{v}$ should not be significant
- The Sargan test tests overidentifying restrictions

A simple example: estimating a demand function

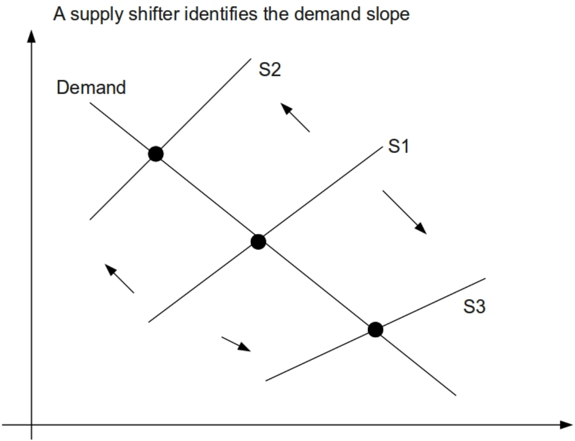
a supply and demand system of equations

- supply function: $q = \gamma_0 + \beta^s p + \gamma x^s + u^s$
- demand function: $q = \alpha_0 + \beta^d p + \alpha x^d + u^d$

At equilibrium, $q = q(x^s, x^d, u^s, u^d)$, $p = p(x^s, x^d, u^s, u^d)$

- Note that $cov(p, u^d) \neq 0$ (OLS is inconsistent)
- "identification" of β^d using a "supply shifter"
  - $cov(x^s, p) \neq 0$(relevance) (because p is a function of x^s)
  - $cov(x^s, u^d) = 0$(exogeneity) (otherwise, x^s is not really a "supply shifter")

A Graphical Interpretation of Identification of Demand



A simple Monte Carlo experiment

$log(wages) = 10 + 0.05 * D + u,\ u \sim N(0,1),\ D = 1$ with prob. 0.3

1. draw N realizations of D
2. draw N realizations of u
3. compute $log(wages)$
4. OLS $log(wages)$ on D and store $\widehat{\beta}_1^r$
5. replicate step 1 to 4 R times
6. examine the empirical distribution of $\widehat{\beta}_1^r$

How do we draw N realizations of D and u?

A random number generator is a device designed to generate a sequence of numbers, called pseudo-random numbers, that appear random

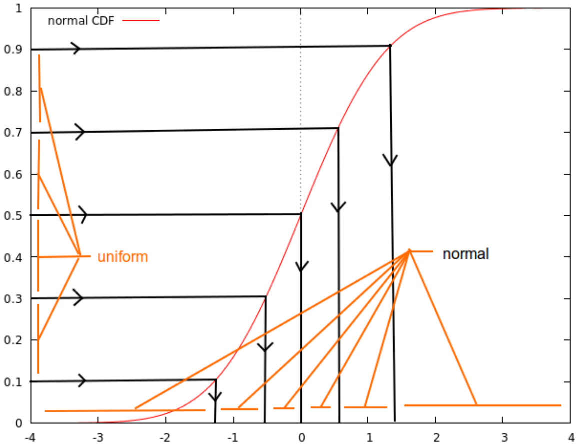
- there are two main methods:
  1. using a physical random phenomenon (i.e. sunspots)
  2. using a computer
- the latter type are determined by a shorter initial number given to the computer, known as "the seed"
- controlling the seed is useful: it permits replication

Pseudo-random numbers of the uniform

- many econometric packages provide pseudo-random numbers from the uniform distribution between 0 and 1
- uniform values between 0 and 1 can be used to generate random numbers of any desired distribution
- how? by passing them through the inverse cdf of the desired distribution

$x \sim N\left(\mu, \sigma^2\right)$

1. generate the uniform $U(0,1)$: u
2. generate the standard normal $N(0,1)$: $z = \Phi^{-1}(u)$
3. compute $x = \mu + \sigma u$


normal CDF
uniform
normal
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4

Multivariate normal pseudo-random numbers

- Any multivariate normal

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

can be expressed as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + A * \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

- A is such that $\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = AA^T$ (Cholesky decomposition)
- $\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$

Random number generation in gretl

Commands to generate random numbers

- `uniform`: draws a series of iid values from the uniform distribution
- `normal`: draws from the uniform distribution
- `genpois`: draws from the poisson distribution
- `randgen`: all purpuse random number generator

We are going to predominantly use `uniform` and `normal`

uniform(#a,#b)

- generates values from the uniform in the interval (a, b)–by default, in the interval (0,1)

Example

- nulldata 500 # "blank" data set with 500 obs.
- set seed 2703 # sets the seed for replicability
- genr x = 100 * uniform(-1,1)

normal($\#\mu$,$\#\sigma$)

- generates values from the normal $N\left(\mu,\sigma^2\right)$–by default, the $N(0,1)$

Example 1

- genr z = normal(5,2)

Example 2: conditional normal distribution

- genr x1 = 20+5*uniform(-1,1)+1.3*normal()
- genr u = uniform(-1,1)+3*normal()
- genr y = 2 + 3 * x1 + 3*u

Sample Covariance of Any Two Variables

- Suppose that we have two random variables, x_1 and x_2, with the following properties

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)$$

- Suppose that we estimate the covariance with samples of size $N = 5, 50, 500, 5000$
- Can we "estimate" the statistical properties of the sample covariance of these two variables for each sample size?
- Can we understand how the asymptotic properties are related with these small sample properties?

Estructura

Objetive: simulate a bivariate normal and estimate the small sample properties of the sample covariance

1. Inicialization: sample size, seed, Cholesky
2. Within the loop:
    1. Simulation
    2. Computation of the sample covariance for different samples
    3. Store results
3. Recovery of results

The gretl Script

```
# *************** Before the loop ********************
nulldata 5000
set seed 547
matrix S = {1,0.5;0.5,1}
matrix A = cholesky(S)

#****** open a loop, to be repeated R=500 times **********
loop 500 --progressive --quiet
 genr u1 = normal()
 genr u2 = normal()
 genr x1 = A[1,1]*u1+A[1,2]*u2
 genr x2 = A[2,1]*u1+A[2,2]*u2
 smpl 5 --random
 genr cov5 = cov(x1,x2)
 smpl full
 smpl 50 --random
 genr cov50 = cov(x1,x2)
 smpl full
 smpl 500 --random
 genr cov500 = cov(x1,x2)
 smpl full
 genr cov5000 = cov(x1,x2)
 store myfirstMC.gtd cov5 cov50 cov500 cov5000
endloop

#*********** we open the results ***********************

open myfirstMC.gtd
summary cov* --simple
```

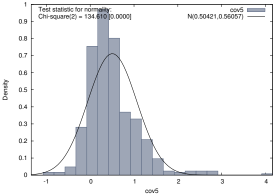
The Monte Carlo Results & the LLN

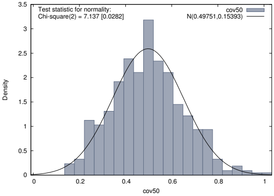
```
Read datafile /home/ricmora/AAOFICIN/CURSOS/MICCUA/materiales/Sesión
3_Tema 1_2_ Propiedades Asintóticas y Simulación en gretl/myfirstMC.gtd
periodicity: 1, maxobs: 500
observations range: 1-500

Listing 5 variables:
  0) const       1) cov5        2) cov50       3) cov500      4) cov5000

? summary cov* --simple

Summary statistics, using the observations 1 - 500

                     Mean         Minimum         Maximum        Std. Dev.
cov5              0.50421        -0.94961          4.0369          0.56057
cov50             0.49751         0.15717          1.0123          0.15393
cov500            0.50006         0.37126         0.63924         0.047988
cov5000           0.50061         0.45830         0.54222         0.015756
```

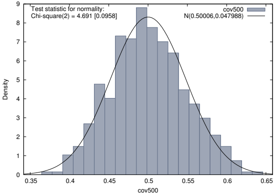
- The average is very similar across different sample sizes, and it is quite close to the population covariance. Why?

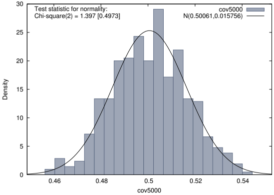
- The standard deviation gets smaller and smaller as the sample size increases. Why?

The Monte Carlo Results & the CLT









Summary

- under classical assumptions, OLS is consistent and asymptotically normal
- when a control is likely correlated with the error term, then OLS is inconsistent
- under general assumptions, 2SLS is consistent and asymptotically normal
- if we want to estimate the price elasticity in a demand equation, we need a "supply shifter"
- within a basic Monte Carlo algorithm we need a random number generator. In `gretl`, this is very easy