Introduction to Programming
Final Exam

Clement Mazet-Sonilhac
SciencesPo Paris

April 2019

Consistency and biais of the OLS estimator: a Monte-Carlo procedure

In this final exam, you will have to:

1. Simulate some data from a data generating process (DGP)
2. Estimate a linear model using generated data
3. Show empirically that the OLS estimator is unbiased
4. Show empirically that the OLS estimator is consistent

Goal. The goal of this exercise is to generate some data from a linear model for which we know the true value of the parameters, then ignore the true values and estimate the parameters via OLS. The idea of this Monte-Carlo procedure is to test whether the OLS estimates are consistent and unbiased. This is a classic exercise in econometrics.

Consider the following data generating process:

$$y_i = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

where the individual variables x_1^i, x_2^i, x_3^i and ε_i follow:

- $x_i^1 \sim \mathcal{N}(2, 0.1)$
- $x_i^2 \sim$ Beta(0.3,0.3)
- $x_i^3 \sim$ U(0,1)
- $\varepsilon_i \sim \mathcal{N}(0, 1)$

and the parameters are given by $\alpha = 2$, $\beta_1 = 0.5$, $\beta_2 = -1$, $\beta_3 = 2$.

Reminder.

- Monte-Carlo simulation: Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle.

- Unbiased estimator: the bias of an estimator is the difference between the estimator's expected value, and the true value of the parameter being estimated, that is
$$Bias(\hat{\beta}) = E(\hat{\beta}) - \beta.$$
- Consistent estimator: a consistent estimator converges in property to the true value of the parameter being estimated when the sample size goes to infinity, that is
$$plim_{n\to\infty}\, \hat{\beta}_n = \beta.$$

Question 1:

Simulate the data for 100, 1000, and 10,000 individuals, and store it in a tibble. The content of this tibble should be similar to the data an econometrician has access to (e.g. the ε_i's and β's are unknown, while the y_i's and x_i's can be observed).

Hint: *You should create a function dgp(s,n) with n being the number of obs. and s the seed (see further). You will need to recompute this dataset for different sample size and different seeds.*

Question 2:

To verify that our dgp() is correct, plot the histogram of $x1$, and plot on top of it the true probability density function of a $\mathcal{N}(2, 0.1)$. Do the same for $x2$.

Question 3:

Estimate via OLS (multivariate linear model, see Lesson 11-12) the following linear model, print the output, and test whether $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are different from zero (separately and jointly).

$$y_i = \alpha + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Question 3 bis (bonus) :

Test for homoskedasticity and, if needed, implement a FGLS correction.

Question 4:

Show that the OLS estimates are unbiased (see reminder).

Hint: *We would like to show that:*

$$Bias(\hat{\beta}) = E(\hat{\beta}) - \beta \approx 0$$

To that aim, we need to approximate $E(\hat{\beta})$. One way to do that on a computer is to simulate many datasets with different seeds (see Monte-Carlo), estimate for each dataset the β's, and compare the distributions of the $\hat{\beta}$'s with the true β's. Since the sample size is finite and the seed is different for each simulation, we'll obtain slightly different data, and therefore different $\hat{\beta}$'s. In practice, you would like to simulate and estimate our model 1000 times, with different seeds each time. Then, plot the distribution $\hat{\beta}_2$, and plot the evolution of our numerical approximation of $E(\hat{\beta}_2)$ as the number of simulation increases.

Cool, but what exactly is a "seed" ? when asking your computer for random numbers, it uses a pseudorandom number generator, i.e. a sequence of numbers whose properties approximate the properties of sequences of random numbers. However, this sequence of numbers is not truly random; it is completely determined by an initial value. This initial value is called the seed. Therefore, by changing the seed, you are ensuring that the random numbers generated are different. In `R`, you can change the seed via `set.seed(<an integer>)`.

Question 5:

Show that the OLS estimates are consistent (see reminder).

Hint: *We would like to show that $\hat{\beta}_1 \to \beta_1$ as $n \to \infty$ In the spirit, it is very similar to Question 4, except that you should iterate on the sample size rather than on the seed. In practice, I would like you to simulate and estimate 10 times our model for sample sizes going from 10 to 100,000. Then, plot the estimates of β_1 against the sample size.*

Question 6 (bonus):

Explain the similarities and the differences between a bootstrapping and a Monte-Carlo procedure.