Advanced Econometrics: Home Assignment 2

2023

Instructions

- The problem set is due November 29, 2023, 23:59. A late submission automatically means 0 points.
- The solutions should be sent to 38341240@fsv.cuni.cz, with the following subject: **'AE_HW_group99_surname1_surname2_surname3'**
- Send me the solution as one Jupyter notebook (.ipynb), named **'AE_HW1_group99_surname1_surname2_surname3.ipynb'** which should contain the main analysis together with commented code. Do not forget to use the full potential of Jupyter notebooks. Show graphs and results as an output of your code. Write the reasoning and other text in markdown cells (which supports headers, Latex equations, pictures, etc.). It is also possible to send me one pdf file with the analysis and one R script (following the naming convention).
- The empirical problems do not necessarily have a unique solution in terms of numbers. You are assessed based on the execution of the analysis not on the right numbers that you should get from the output. The emphasis is put mainly on meaningful presentation and the extent of your knowledge.
- Please, use 'set.seed()' function, so I can replicate your results.
- If you have any questions concerning the homework, do contact me by mail. Do it rather sooner than later.

Problem 1: GMM (3 points)

In the dataset **hw_data.csv**, you have a time series which comes from the Moving Average process with q lags, MA(q) process. The MA(q) process is defined as follows:

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

$$\epsilon_t \sim iidN(0, \sigma^2)$$

You will need libraries 'gmm' and 'tseries' to answer the following questions. Note that if you do not answer correctly point a), all consecutive questions will be wrong, hence no points can be earned.

(a) How would you identify the lag q? Identify the lag q.

(b) Derive the moment conditions function for the process you identified previously and write a corresponding function in R. Use more moment conditions than the number of coefficients that you want to estimate.

(c) Estimate the model using 'gmm' with an identity weighting matrix. Provide the output and interpret the coefficient significance and the J-test statistic.

(d) Estimate the model using 'gmm' with an optimal weighting matrix. Provide the output and interpret the coefficient significance and the J-test statistic.

(e) Compare the results from c) and d).

Problem 2: Delta method (2 points)

Let us model the variables X_1 and X_2 as follows:

$$X_1 = 0.2V_1 + 0.7V_2 + 0.3V_3$$

$$X_2 = 4V_2 - 3V_3$$

$$V_1 \sim N(0,1), V_2 \sim \chi^2(3), V_3 \sim U[1,2]$$

The linear model is defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Let us assume, that we are interested in quantity Φ:

$$\Phi = ln(\beta_0 + \beta_1^2) - (\beta_1^{\frac{1}{2}})$$

Estimate the value of Φ and also the variance of Φ, $Var(\Phi)$. To estimate $Var(\Phi)$ use the next methods:

(a) Delta method (by hand)

(b) Delta method (using 'deltamethod' function)

(c) Bootstrap (with 1000, 10000 replications)

(d) Compare the results.

Problem 3: Bootstrap (2 points)

Assume a random sample from Exponential distribution. $X \sim exp(\lambda)$.

(a) Generate 200 observations from the exponential distribution for $\lambda = 2$.

(b) Check the sample mean and variance and compare them to the theoretical values.

(c) Plot the histogram of generated data, kernel density approximation, theoretical density of exponential distribution and exponential Q-Q plot. Discuss.

(d) Use brute force bootstrapping to obtain the Bootstrap mean, standard errors and bias-reduced estimate of the mean. Use 100,000 bootstrap replications.

(e) Now, use the 'boot' function with 100,000 bootstrap replications.

(f) Obtain the confidence intervals, use different types and compare the results. Which type is appropriate to use in this case?

(g) Plot the bootstrapped mean. Discuss.

Problem 4: Endogeneity (2 points)

Let us follow the idea of the first exercise from Seminar 6 but for now, we create another artificial dataset containing 250 observations:

$$z_1 \sim N(3, 9), z_2 \sim N(2, 0.5^2), z_3 \sim N(0, 4), z_4 \sim N(2.8, 0.2^2)$$

$$\epsilon_1 \sim N(0, 1.2^2), \epsilon_2 \sim N(0, 1.2^2), \epsilon_3 \sim N(0, 1.2^2)$$

$$x_1 = 0.8z_1 + 2z_2 - 3z_4 + 0.5\epsilon_1$$

$$x_2 = 0.75z_2 - z_4 + 0.5\epsilon_2$$

$$x_3 \sim N(0, 1)$$

$$y = 1 + 2x_1 - 0.5x_2 + 2.5x_3 + \epsilon_3$$

Using the dataset, we should estimate the following model:

$$y = \alpha + \beta_1 x_1 + \beta_3 x_3 + \epsilon$$

(a) Discuss the nature of the endogeneity problem in the system above. You might check important correlations and you should explain the difference between x_1 and x_3. Do you expect to observe any bias within the OLS estimation? Explain why and discuss the source of possible bias. What solutions do you suggest?

(b) Estimate the model by OLS and interpret.

(c) The data set includes some potential IV candidates: z_1; z_2; z_3; z_4. What assumptions need to be satisfied to have a 'good' instrument? Which of these candidates seem to be 'good' instruments and why? Test their relevancy statistically. Is there any invalid, irrelevant, or weak instrument?

(d) Based on section c), choose the best instrument and run the IV regression. Run also 2SLS regression using all 'good' instruments. Compare coefficient estimates and standard errors.

(e) Finally, test for the endogeneity using the Hausman test. Report and interpret the results.

(f) Using an extended dataset (simulating more data from the data generating process), show that OLS is not a consistent estimator of β_1 and β_3. Show that 2SLS provides consistent results.