Advanced Econometrics 2020 - Homework 2

Deadline: 23:59:59, Friday November 27, 2020

Instructions[1]:

- **Submit by e-mail to `matej.nevrla@fsv.cuni.cz` with the following subject: "AdvEcox HW2 2020: Group surname1, surname2, surname3 "**.
- Form groups of three yourself.
- As a solution, provide **one Jupyter Notebook** with R source-code. Code should be properly commented, interpretations of results as well as theoretical derivations[2] should be written in markdown cells. This is the only file you need to send.
- Use "set.seed()" function, so I can replicate your results.
- Be concise (no lengthy essays please). Although, be sure to include all important things as I cannot second-guess your work.
- **The problem set is due on 27th November. Late submission automatically means 0 points.**
- If you have any questions concerning the homework, do contact me by mail. Do it rather sooner than later.

[1] The contact person for this homework is Matěj Nevrla, the same mail as for submission of homeworks.

[2] If you prefer not to write formulas in LaTeX, you can send PDF with your derivations and interpretations in additional file and R code in Jupyter Notebook.

Problem 1 Bootstrap

(1.5 points)

Consider dataset `city` from the package **boot**. Data consists of random sample of size $n = 10$ from the population of big cities in the US and each observation corresponds to the number of inhabitants of a given city (in thousands) in year 1920 (variable u) and number of inhabitants of a given city in year 1930 (variable x).

Let (U, X) be a random vector corresponding to the number of inhabitants in the years 1920 and 1930, and (u_i, x_i), $i = 1, \ldots, 10$, be our random sample.

Consider the following ratio

$$R \equiv \frac{E(X)}{E(U)}.$$

a) Draw the dependence between x_i and u_i using the dataset.

b) Estimate the value of R from the random sample.

c) Compute the bootstrap standard error and the bootstrap bias of the estimate of R. Use both "brute force" bootstrapping and function `boot` from the respective package.

d) Compute the 95% confidence intervals using both percentile and normal approximation method. Argue whether the normal approximation is valid in this case.

e) Test (without refinement) H_0 that $R = 1$ against the alternative $H_A : R \neq 1$. Calculate bootstrap t-statistic and compare it with critical values of the standard normal distribution. Use also the percentile method for testing H_0 and compare the results.

Problem 2 Endogeneity

(2 points)

Let us follow the idea of the first exercise from Seminar 6 but for now we create another artificial dataset containing 300 observations *(note that although variance of RVs is specified below, R commands often require to specify standard deviation instead)*:

$$
\begin{aligned}
z_1 &\sim N(2,3^2), \quad z_2 \sim N(2,1.5^2), \quad z_3 \sim N(0,2^2), z_4 \sim N(1.8,2.5^2) \\
\epsilon_1 &\sim N(0,1.5^2), \epsilon_2 \sim N(0,1.5^2), \epsilon_3 \sim N(0,1.5^2) \\
x_1 &= 0.3z_1 - z_2 + 0.9z_4 + 0.75\epsilon_1 \\
x_2 &= 0.75z_2 + 0.75\epsilon_2 \\
x_3 &\sim N(0,1) \\
y &= 1 + 2.5x_1 - x_2 + 0.45x_3 + \epsilon_3
\end{aligned}
$$

On the dataset, we should estimate the following model:

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

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.

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 in order 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 d), 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 extended dataset (simulate more data from the data generating process), show that OLS is not consistent estimator of β_1 and β_3. Show that 2SLS provides consistent results.

Problem 3 GMM

(1.5 points)

In the dataset `hw_data.csv`, you have a time series which comes from Moving Average process with q lags - MA(q) process. 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? What is the lag?

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

c) Estimate the model using `gmm` and both identity and optimal weighting matrix. Provide the output and interpret the coefficient significance and the *J-test* statistics.

Problem 4 CAPM beta of Apple Inc.

(3 points)

You are going to estimate CAPM betas for Apple Inc. on two different time periods and compare the results. Inference on the estimated parameters will be performed using bootstrap.

CAPM beta for company i can be estimated by regressing its returns on market returns using simple OLS

$$\begin{aligned} r_t^i - r_t^f &= \alpha_i + \beta_i (r_t^M - r_t^f) + \epsilon_t^i \\ r_t^{e,i} &= \alpha_i + \beta_i r_t^{e,M} + \epsilon_t^i \end{aligned}$$

where r_t^i ($r_t^{e,i}$) is log (excess) return of company i, r_t^M ($r_t^{e,M}$) is log (excess) market return, and r_t^f is risk-free interest rate, all at time t.

1. Download adjusted daily prices of Apple Inc. (ticker `'aapl'`) for year 2008 and 2017, separately. Download daily data for Nasdaq Composite Index (ticker `'^ixic'`), which will be used as a proxy variable for market. Chicago Board Options Exchange (CBOE) 10y interest rate T-note (ticker '`^tnx`') will be used as a risk-free rate.

2. Compute log-returns for all series.

3. Draw histogram and kernel density approximation of Apple log returns. Compare it with normal distribution with the same mean and standard deviation. Does samples considerably deviate from normality? Which noticeable features do you observe in the data? Perform this whole analysis on both time periods separately.

4. Compute CAPM beta for Apple using simple OLS on each time period separately. Compare the obtained results.

5. Use the nonparametric bootstrap to compute bootstrap standard error of CAPM beta estimate based on 1000 bootstrap replications and bootstrap sample size equal to the size of the original sample. Use `boot` command. Perform this whole analysis on both time periods separately.

6. Graphically compare histogram and kernel density approximation (use gaussian kernel only) of the boot-based set of 1000 bootstrapped CAPM betas with normal distribution. Compute 95% percent confidence intervals for CAPM beta for both time periods using percentile method and method based on normal approximation. Which method is in this case probably more suitable. Based on the estimated confidence intervals, do the estimates of CAPM beta significantly differ over these 2 periods?

7. Another approach how to use bootstrap is instead of resampling the observations is to resample the estimated residuals. Estimate CAPM beta by OLS, save the estimated residuals, and perform bootstrap by resampling the residuals and generating new values of dependent variables as

$$r_t^{e*} = \alpha_i + \beta_i r_t^{e,M} + \epsilon_t^{i*} \tag{1}$$

where ϵ_t^{i*} are resampled from empirical distribution function. In each bootstrap replication, calculate values of dependent variable for all values of independent variable. The rest of the procedure is the same as in the case of resampling the observations. Perform this analysis on one time period of your choice. Compare the results obtained from this approach and from the traditional approach from above. Why this method may not be the optimal in this case?