Advanced Econometrics 2019 - Homework 2

Deadline: 23:59:59, Friday November 29, 2019

Instructions[1]:

- **Submit by e-mail to `matej.nevrla@fsv.cuni.cz` with the following subject: "AdvEcox HW2 2019: Group surname1, surname2, surname3 "**.
- Form groups of three yourself.
- As a solution, provide 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 29th November. Late submission automatically means 0 points.**
- If you have any questions concerning the homework, do contact me by mail and we can set up a consultation. Do it rather sooner than later, I won't give any consultation concerning the homework after the 27th November.

[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 `scor` from the package **bootstrap**. Data consists of student's performance in 5 subjects: mechanics (column `mec`), vectors (`vec`), algebra (`alg`), analysis (`ana`), and statistics (`sta`). The size of this sample is $n = 88$.

a) Consider variables `alg` and `sta`. Estimate the correlation coefficient between these two variables. Compute standard deviation, bias, and 95% confidence interval (use various methods) of this estimate using bootstrap. Use function `boot` and at least 1000 bootstrap replications.

b) Estimate the covariance matrix of the dataset. The ratio between the biggest eigenvalue and sum of all the eigenvalues of the covariance matrix corresponds to the proportion of variability of the original variables explained by the first principal component. Estimate this quantity. Compute standard deviation, bias, and 95% confidence interval (use various methods) of this estimate using bootstrap. Use function `boot` and at least 1000 bootstrap replications.

Problem 2 **Endogeneity**
(2 points)

Let us follow the idea of the first exercise from Seminar VI 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)*:

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

On the dataset, we should estimate the following model:

$$y = const + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

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_2 and x_3. Explain why do you expect to observe any bias within the OLS estimation.

b) Estimate the model by OLS and interpret. Using increasing sample size, *show that the OLS is not consistent estimator*. To do that, simulate 5000 observations from the data generating process and estimate the model on the increasing number of observations.

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 3., choose the best instrument and run the IV regression. Run also 2SLS regression using all 'good' instruments. Compare coefficient estimates and standard errors

e) Test for the endogeneity using the Hausman test. Report and interpret the results.

Problem 3 **GMM**
(1.5 points)

The mean of χ^2 random variable with d degrees of freedom is $\mathbb{E}(Y) = d$ and the variance is $\mathbb{V}ar(Y) = 2d$.

a) Based on the two moment conditions mentioned above, specify 2 sample moment conditions which can be utilized in the GMM estimation.

b) Simulate random sample of length 500 and $d = 1$. Use the 2 sample moment conditions to estimate d using GMM using both identity and optimal weighting matrix.

c) Use the $J - test$ and verify whether the moment conditions are satisfied.

d) Based on 1000 simulation of length 500, compare the efficiency of the following estimators

   (a) MM estimator using the first condition,

   (b) GMM estimator using both conditions and identity weighting matrix,

   (c) GMM estimator using both conditions and optimal weighting matrix,

   (d) MLE estimator.

   Moreover, plot the simulated distribution of each estimator.