AMERICAN UNIVERSITY - CENTRAL ASIA 25.11.2022
Econometrics I (ECO 608)
Nurgul Tilenbaeva

Homework Assignment 2

For this homework assignment that comprises ***10% of your final grade*** *you will need to work in groups of two or three people. Each group will need to hand in one solution: one member of the group will need to send the following materials by email to* `tilenbaeva_n@auca.kg` *and put his/her group-mates in the copy of the email:*

- *Pdf file with screenshots of Stata outputs, interpretation of results, answers to the questions, etc.*
- *Do-file containing commands used to answer the questions.*

Please make sure to indicate the name(s) of your group members in both documents. Please note that all the solutions need to be typed, so please allow enough time for typing the solutions out. No handwritten solutions will be accepted.

Deadline *for submission is:* ***4 December 2022, 23:59****. If you submit your assignment after the deadline, the maximum you can get for your work will be the following:*

- *1-5 minutes late: 8%*
- *6-59 minutes late: 7%*
- *60 minutes-24 hours late: 5%*
- *after 24 hours: 0%*

1 Specification Error

(51 points)

1. Import the Stata data file "`ceosal1`" from the e-course platform.
2. **(2 points)** Run the linear regression for the model explaining CEO salary, given by

   $log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 roe + \beta_3 rosneg + u$

where

`salary`=CEO salary in thousands of US dollars;

`sales`=firm sales, in millions of US dollars;

`roe`=return on equity;

`rosneg`=a dummy variable, which is equal to "1" if the return on firm's stock `ros` < 0, and "0" if `ros` ≥ 0.

3. **(10 points)** Apply a RESET test manually by running the corresponding regression using the fitted values from estimating the model in part (2), and performing hypothesis-testing by hand at `5% significance level`. Make sure to add the squared, cubed and fourth-power terms. Is there evidence of functional form misspecification in the equation?

4. **(3 points)** Now check your results by running a RESET test in Stata.

5. Import the Stata data file `"ceosal2"` from the e-course platform.

6. **(15 points)** Now we would like to test the model

   $log(salary) = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u$

   against the model

   $log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 log(mktval) + \beta_3 log(ceoten) + u$

   and vice versa.

   `salary`=CEO salary in thousands of US dollars;

   `sales`=firm sales, in millions of US dollars;

   `mktval`=firm market value, in millions of US dollars;

   `ceoten`=years as CEO with company.

   Apply a test using the approach by Mizon and Richard (1986) manually, by running the corresponding regression, and performing hypothesis-testing by hand at `5% significance level`. Do not forget to perform two tests, as we are testing two models against each other. Based on your results, which model you would prefer?

7. **(6 points)** Check your results by running an F test in Stata.

8. **(15 points)** Now use the Davidson-MacKinnon test manually to choose among the models in 1.6., by running the corresponding regressions

using the fitted values, and performing hypothesis-testing by hand at `5% significance level`. Do not forget to perform two tests, as we are testing two models against each other. Based on your results, which model you would prefer?

2 Omitted Variables

(18 points)

1. Import the Stata data file `"caschool"` from the e-course platform. This data set contains information on test performance, school characteristics and student demographic backgrounds for California school districts.

2. **(2 points)** Estimate the regression model by OLS

   $testscr = \beta_0 + \beta_1 str + u$

   where

   `testscr`=average test score in a school district;

   `str`=student to teacher ratio;

3. **(5 points)** We are worried that there could be an omitted variable, namely the percentage of English learners in the school district: it is plausible that the ability to speak, read and write English is an important factor for successful learning. Therefore, students that are still learning English are likely to perform worse in tests than native speakers. Also, it is possible that the share of English learning students is bigger in school districts where class sizes are relatively large (i.e., student to teacher ratio is high): think of poor urban districts where a lot of immigrants live. Given this information, what is the expected direction of the bias on the coefficient for `str` in 2.2? Explain.

4. **(5 points)** Add the variable `el_pct`, which denotes the share of English learning students in the school district, to the regression to account for omitted ability bias. What happens to the estimated coefficient on `str`? Is the coefficient on `el_pct` statistically significant? Are the results in line with your prediction in part 2.3?

5. **(6 points)** What are the rules of thumb for including variables in the model?