General instructions for the Exam

Advanced Econometrics I

February 23th, 16:00 h

- A printout of the problem set (Klausur) and some blank pages of paper will be made available at the beginning of the exam.
- Please make available your responses (solutions) of the cases via Word, Open-Office or some other text editor. You may copy the output of programs like R, Gretl or Stata in your text editor. Please save the complete content of the text editor as a PDF-File.
- If necessary, you can also submit a handwritten solution on the paper that was distributed at the beginning of the exam.
- Start your text file with indicating your name and your student ID.
- The name of your PDF-file should be just your student ID (no names).
- Please upload the PDF-File with your solution to the Ilias-Folder “Submitted files” 60–70 minutes after the begin of the exam. You cannot assume that the exam will be accepted if you upload it later than 70 minutes.
- If you encounter technical problems during the uploading process, you may also submit your PDF-file via email attachment. But you need to make sure that you send the email (see next bullet point) within the time limit of 70 minutes.
- Good luck!

Case 1: Cross-section regression

The Gretl data set `happiness.csv` contains data of 2015 from the German Social-Economic Panel (GSOEP) provided from the DIW.

hinc: Household income

unemployed: Dummy variable: 1 = unemployed

age: person's age

married: Dummy variable: 1 = married

migback: Dummy variable: 1 = with migration background

health: self assessment (scale: 0 – 10) for the person's health

happiness: self assessment (scale: 0 – 10) for the person's happiness

Tasks

Create a new variable that results from multiplying `happiness` with the factor = `MatNo`/7123456, where `MatNo` is your 7-digit student ID. Use this transformed variable (say `happiness_new`) instead of the original variable.

12 a) Run an OLS regression of `happiness_new` on the remaining variables. Is the error homoskedastic and normally distributed? How do you adjust the estimation if you find heteroskedasticity? What are the implications for the OLS estimator if the errors are NOT normally distributed?

10 b) How much household income needs to be increased in order to raise `happiness_new` by one unit (holding all other variable constant). Compute also the standard error for your estimate.

10 c) Include the square of `age` in the regression and test the hypothesis that age does not affect happiness. At what age happiness is minimized? Test the null hypothesis that happiness achieves the minimum at `age` = 40 years.

10 d) Why is it problematical to assess the importance of some variable for explaining the dependent variable by its coefficient β? In practice the importance of the variable is often measured by the t-statistic or the beta-coefficients. Order the importance of the regressors by using (i) the t-statistics and (ii) the beta-coefficients.

8 e) Some colleague suggests to use the logarithm of `happiness_new` as the dependent variable. You observe that such a regression produces a smaller R^2 than the regression with the `happiness_new` variable. Does that mean that the original regression performs better than using the dependent variable in logarithms?

Binary response model

.

To investigate the probability to be an owner of a house, the dataset `house_owner.csv` with information on 8993 individuals is available. The variable descriptions are as follows:

income: net income, classified from 1 (lowest category) to 9 (largest category)

age: age classified from 1 (below 18) to 7 (above 65)

educ: classified from 1: less than 9 years of schooling to 6: academic degree

hh_size: number of persons living in the household

ethnic: 1 = White American, 0 = other ethnics (African, Hispanic, Asian etc.)

own_house: 1 = home owner, 0 = no home owner

x: Computer generated random variable not related to all other variables

Tasks

10 a) Estimate an OLS regression of `own_house` on all remaining variables (including `x`). For which distribution of the errors in the latent regression model this estimation method provides consistent estimates? Explain why you need robust standard errors for computing valid t-statistics.

10 b) Estimate a Probit model and compute the (McFadden) pseudo-R^2 for the model. Compute also the McFadden R^2 for the linear probability model in a). Which model provides the better fit? Why are the estimated coefficients so different whereas the t-statistics are fairly similar?

10 c) How does the probability for owning a house changes in the two models estmated in a) and b) if the household size increases by one person (and all other variables remain the same)? Is the marginal probability effect scale-invariant?

10 d) Assume that it is known that the intercept in the latent model is $\beta_1 = -1$. Compute the estimated threshold level that applies when transforming the latent variable y_i^* into the observed variable y_i?

10 e) Compute the 4-cell matrix for the number of actual and predicted outcome based on a threshold level of 0.5. Compute the error rates for predicting the outcome `own_house`=0 and `own_house`=1 (that is the relative frequencies of the wrong predictions). Do you think that the model does a good job in predicting the outcome? What can you do to improve the performance?