Advanced Econometrics: Homework 1

2023

Instructions

- Form groups of **three** yourself. Write down your names in the following spreadsheet:
  `https://docs.google.com/spreadsheets/d/1P7jAIqCtRR0cUTuqlS6vSCLWc--8JrL-e-vkal5S0_Y/edit?usp=sharing`
- The problem set is due by 23:59 **November 8, 2023**. A late submission automatically means 0 points.
- The solutions should be sent to lenka.nechvatalova@fsv.cuni.cz,
  with the following subject:
  'AE_HW1_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 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 the meaningful presentation and the extent of your knowledge.
- If you have any questions concerning the homework, please discuss them with your teammates first. You can contact me by mail, and we can schedule a consultation if needed. Do it sooner rather than later; I won't give any consultation concerning the homework after November 6.

Problem 1

Simulate hypothetical data about used car prices according to the following equation

$$price_i = 1000 - 0.0005 * milage_i^2 - 5 * years\_old_i + 50 * convertible_i + 100 * luxury\_brand_i + \epsilon_i,$$

where $milage_i \sim N(1000, 300^2)$, $years\_old_i \in \{2, ..., 15\}, convertible_i \in \{1, 0\}, luxury \in \{0, 1\}$ and the error term has distribution $\epsilon_i \sim N(0, 15)$. Choose your own size of the sample $i = 1, ..., N$. (Note that the bigger sample the more precise estimates will be obtained. So the interesting case is rather to take relatively low number of observations.)

1. Describe the data using descriptive statistics and figures.
2. Estimate the regression model as

   $$price_i = \alpha_0 + \alpha_1 * milage_i + \alpha_2 * years\_old_i + \alpha_3 * convertible_i + \alpha_4 * luxury\_brand_i + \epsilon_i.$$

   What is wrong? Are some of the estimates biased? If not, why is it so, what concept would you use to prove that?
3. Estimate the regression model with the *true* functional specification. Test whether the estimates are equal to the true values. Interpret the test results.

4. Simulate the prices with the following change:

   $price_i = 1000 - 0.0005 * milage_i^2 - 5 * years\_old_i + 50 * convertible_i + 100 * luxury\_brand_i + milage_i * \epsilon_i$.

   Now you have data that contain error that is proportional to *milage*. What is the issue that you would run into, if you ran regression model as in part 3? How would you test formally for this problem? Perform the tests. Compute standard errors that account for such problem. How would you correct your specification, given that you know the exact form of it?

5. Run quantile regression for quantiles from 5% to 95% by 5% and plot the results. Is it useful? Does it bring any new information? Why, why not?

Problem 2

You are provided data 'problem_2_dataset.Rdata'. The dataset contains the results of World Banks 1997 Vietnam Living Standards Survey. The following variables in the dataset are of interest:

- *lhhexp1:* log of total expenditure over the year,
- *lhhex12m:* log of expenditures on health care,
- *farm:* dummy for living on a farm,
- *urban98:* dummy for living in an urban area,
- *age:* age of respondent,
- *sex:* sex of respondent (1: male, 2: female).

Do the following:

(a) Describe the data statistically. Make some descriptive plots.

(b) Fit a linear model with *lhhex12m* as a dependent variable and the rest of the variables as regressors; include constant. Describe and interpret the results.

(c) Fit quantile regression of the same model specification for quantiles from 5% to 95% by 10% steps and plot the results. Interpret the results. Discuss what information did you get from the quantile regression.

(d) With 95% confidence level for which quantiles are your coefficients different from OLS coefficients? (Do this visually from the figures; the results are obvious from them.) Contrast the linear and quantile regression results.

(e) We know that quantile regression does not work by considering only some portion of the data (i.e. one decile). Demonstrate (if it is indeed the case) that splitting the data into deciles (based on the response variable) and performing linear regression for each decile separately (incorrect method) leads to different "decile" betas compared to the (correct) betas from quantile regression. For simplicity, consider only one independent variable, *lhhexp1* and intercept. Plot betas from both methods the same way as in the case of quantile regression (confidence intervals are not necessary). Reason why there is a difference between the two methods.

Problem 3

Weibull distribution has the following density:

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, x \geq 0, \alpha, \beta > 0$$

(a) obtain log-likelihood function for a random sample of n observations

(b) obtain the first order condition equations for α and β. Note that we get the solution for α in terms of data x and β. While for the second equation, we get only an implicit solution for β. How would you continue to obtain the maximum likelihood estimators? (it is possible to provide a scan of handwritten solution in your ipynb)

(c) obtain a Hessian matrix of the log-likelihood function.

You are provided with sample data generated by the Weibull distribution in 'problem_3_dataset.Rdata'.

(d) obtain maximum likelihood estimates of α and β, and estimate the asymptotic covariance matrix for the estimates.

(e) obtain maximum likelihood estimate of α under the hypothesis that $\beta = 1$.

(f) carry out Wald test with $H_0 : \beta = 1$

(g) carry out likelihood ratio test with $H_0 : \beta = 1$

(h) carry out Lagrange multiplier test with $H_0 : \beta = 1$