LMAT 1271 - Project 2021-2022

Objectives

- This project is divided into two main exercises :
  1. The study and comparison of point estimation procedures for the unknown parameter of a distribution.
  2. A short regression analysis.
- More globally, a central objective consists in using the software R in order to illustrate various concepts of the 'way-of-thinking' of a statistician.

Instructions

- This project is **mandatory** and counts for **4 points** out of the final note on 20 points for the course.
- This project has to be done in **groups of 3 to 5 students** (exceptions allowed by informing us).
- Don't exchange your work/code with other groups. Reports that are too similar will get a 0.
- **Important** : don't wait until the last week to start this project ! The first part of the project can be done completely during the Easter holiday, and this will help you for the end of the course.

Report contents

- Your report may be in **English or in French**. It needs to start with a cover page specifying the first and last names and the NOMAs of all group members. It needs to end with an Appendix containing :
  1. The plots asked in the project.
  2. Your R code.

  The body of the work is not limited in number of pages.
- Grades are granted to the members whose names are on the PDF.
- The clarity and conciseness of your analysis and graphs are very important.
- The absence of R code will lead to a lower grade.

Report submission

- Every group needs to submit their work in a single **PDF** file **before May 20, 2022 at midnight**. Submission after the deadline will not be accepted.
- To submit your report, send it by e-mail to hortense.doms@uclouvain.be and rainer.vonsachs@uclouvain.be. You will receive a notification once your report has been received.

Part 1 - Point estimation

Context of the exercise

Your engineering team just landed a consulting contract with a company interested in the electricity consumption of its machines. In a first part, you are asked to determine how electricity consumption is evenly distributed across the different machines of the same type. To this end, you use the Gini coefficient. In a nutshell, it is an index ranging from 0 to 1 measuring the inequality featured in a distribution. A value of 0 denotes that all your machines use the same amount of electricity while a value of 1 means that all the electricity is used by a single machine.

We assume that all of the n machines operate independently and their daily electricity consumption (in MWh) can be modelled as a random variable X with the following density function

$$f_{\theta_1,\theta_2}(x) = \begin{cases} \frac{\theta_1 \theta_2^{\theta_1}}{x^{\theta_1+1}} & \text{if } x \geq \theta_2 \\ 0 & \text{otherwise,} \end{cases}$$

with $\theta_1 > 2$ and $\theta_2 > 0$.

Definitions

— Quantile function. The quantile function of X evaluated at t, denoted by $Q(t)$, is the value x_t which solves $P(X \leq x_t) = t$.

— The Gini coefficient. The Gini coefficient of X is defined as

$$G = 2 \int_0^1 p - \frac{\int_0^p Q(t)dt}{E[X]} dp,$$

where $E[X]$ denotes the expected value of X.

Results and tools

— Inverse transform sampling. If X is a continuous random variable with quantile function $Q(.)$ and $U \sim U[0,1]$, then $Q(U)$ and X have the same distribution.

Questions

(a) Derive the quantile function of X (Hint : You should not bother about the part for which $x < \theta_2$ and should simply set $Q(0) = 0$.) We will denote it by $Q_{\theta_1,\theta_2}(.)$.

(b) Derive the Gini coefficient of X. We will denote it by G_{θ_1,θ_2}.

(c) Derive the maximum likelihood estimator of G_{θ_1,θ_2}. Call this estimator $\hat{G}_{\text{MLE}}$.

(d) Propose a method of moment estimator for G_{θ_1,θ_2}. Call this estimator $\hat{G}_{\text{MME}}$.

(e) Set $\theta_1^0 = 3$ and $\theta_2^0 = 1$. Generate an i.i.d sample of size $n = 20$ from the density $f_{\theta_1^0,\theta_2^0}$. In order to achieve this, you can make use of the inverse transform sampling. Using this sample, compute $\hat{G}_{\text{MLE}}$ and $\hat{G}_{\text{MME}}$.

(f) Repeat this data generating process $N = 1000$ times (with the same sample size $n = 20$ and the same (θ_1^0, θ_2^0)) Hence, you obtain a sample of size N of each estimator of G_{θ_1,θ_2}. Make a histogram and a boxplot of these two samples. What can you conclude ?

(g) Use the samples obtained in (f) to estimate the bias, the variance and the mean squared error of both estimators What can you conclude ?

(h) Repeat the calculations in (f) for $n = 20, 40, 60, 80, 100, 150, 200, 300, 400, 500$. Compare the biases, the variances and the mean squared errors of both estimators graphically (make a separate plot for each quantity as a function of n). What can you conclude ? Which estimator is the best ? Justify your answer.

(i) Create an histogram for $\sqrt{n}(\hat{G}_{\text{MLE}} - G_{\theta_1^0,\theta_2^0})$, for $n = 20$, $n = 100$ and $n = 500$. What can you conclude ?

NB : This part is the larger part, it can be done after the TP8 (as of April 1, 2022).

Part 2 : Regression

Context

In the second part of the project, the company wants to understand how electricity consumption is linked to productivity (You can interpret productivity as the daily amount in 1000 Euros that the company gains when the machine operates). You gather a dataset made of 40 independent observations for which you observe the following variables (you can find this dataset on Moodle).

Name of the variable	Description
Y	Productivity in thousands of Euros per day
X	Electricity consumption in MWh

Questions

(a) Is it reasonable to fit a linear regression model between productivity (Y) and electricity consumption (X) ? If no, what transformation of X and/or Y would you propose to retrieve a linear model ? Justify your answers. (Hint : Graphical representation may help you visualize how the variables - including the residuals - behave). For the rest of the exercise, you should work with the transformed variables, which we refer to as X^* and Y^* (Hint : Note that it may be that $Y = Y^*$ and/or $X = X^*$.) Write down the obtained model .

(b) Mathematically derive the marginal impact of X on Y in your model . This is computed via the following formula : $\frac{\partial E[Y|X=x]}{\partial x}$. Provide interpretations .

(c) Is the linear effect significant ? Choose the adequate test for testing linear significance. Compute the p-value of this test . Based on the resulting p-value, what can we conclude ? Analyse the value of the linear effect .

NB : This part can be done after TP12 (latest).