LMAT 1271 - Project 2022-2023

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.
- Please add the function `set.seed(xxx)` to your code with xxx being a three digit number of your choice. Make sure you run this function when you generate your results. This will allow the examiner to generate exactly the same results as you.
- **Important** : don't wait until the last week to start this project ! The first part of the project can be done immediately after 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

By May 26, 2023 at midnight, each group must to submit their results by sending the following 2 documents by e-mail to hortense.doms@uclouvain.be and rainer.vonsachs@uclouvain.be :

1. A single PDF file containing the report and the appendix.
2. A single .R file containing your R code enabling to reproduce the claimed results. Its subdivision must follow the same structure as in the questionnaire.

You will receive a notification once your report has been received. Submission after the deadline will not be accepted.

Part 1 - Point estimation

Part 1.1. : Context and preparatory analysis

Let the density function of a random variable T be

$$f_{\alpha,\beta}(t) = \alpha\beta t^{\alpha-1}(1-t^{\alpha})^{\beta-1} I(0<t<1)$$

where $\alpha > 0$ and $\beta > 0$ are unknown parameters.

(a) Calculate q_p, the quantile of order p of this population.

(b) Calculate $E(T^k)$ where k is a positive constant. *Hint :* $\int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ where $\Gamma(a) = \int_0^{+\infty} t^{a-1}\exp(-t)dt$.

NB : This part can be done immediately.

Part 1.2 : Estimation

Let $T_1, T_2, \ldots, T_n$ be an *iid* random sample from a population with density function (the same as in Part 1.1)

$$f_{\alpha,\beta}(t) = \alpha\beta t^{\alpha-1}(1-t^{\alpha})^{\beta-1} I(0<t<1)$$

where $\alpha > 0$ and $\beta > 0$ are unknown parameters. The goal of this part is to construct three different estimators for the quantile of order 0.95 of this population using the *iid* sample $T_1, T_2, \ldots, T_n$.

(a) Without doing any calculations, propose a first estimator of $q_{0.95}$ based on the ordered statistic of the original sample $T_1, T_2, \ldots, T_n$. We will denote this estimator by $\widehat{q}_S$. Justify your answer.

(b) Show that $(\widehat{\alpha}_M, \widehat{\beta}_M)$, where $\widehat{\alpha}_M$ and $\widehat{\beta}_M$ satisfies the equations

$$\left(\sum_{i=1}^n T_i^{2\widehat{\alpha}_M}\right)\left(n+\sum_{i=1}^n T_i^{\widehat{\alpha}_M}\right)\left(\sum_{i=1}^n T_i^{\widehat{\alpha}_M}\right)^{-2} = 2 \quad \text{and} \quad \widehat{\beta}_M = \frac{n}{\sum_{i=1}^n T_i^{\widehat{\alpha}_M}} - 1\ ,$$

is an estimator of (α, β) obtained by the method of moments.

(c) Show that the maximum likelihood estimator $(\widehat{\alpha}_L, \widehat{\beta}_L)$ of (α, β) satisfies the equations

$$\frac{1}{\widehat{\alpha}_L} + \left(\frac{1}{\sum_{i=1}^n \ln(1-T_i^{\widehat{\alpha}_L})} + \frac{1}{n}\right)\sum_{i=1}^n \frac{T_i^{\widehat{\alpha}_L}\ln T_i}{1-T_i^{\widehat{\alpha}_L}} = -\frac{1}{n}\sum_{i=1}^n \ln T_i$$

and

$$\widehat{\beta}_L = \frac{-n}{\sum_{i=1}^n \ln(1-T_i^{\widehat{\alpha}_L})}$$

(d) Deduce from (b) and (c) two other estimators of $q_{0.95}$. Justify your answer. We will denote these two estimators respectively by $\widehat{q}_M$ and $\widehat{q}_L$.

Part 1.3 : Simulations

The goal of this third part is to perform simulations in order to compare the performance of the three different estimators developed in Part 1.2 and finally conclude which one the best. First, let α_0 and β_0 be respectively values of α and β of your choice (do not take $\beta = 1$ and $\alpha = 1$).

(a) Generate an iid sample of size $n = 20$ from the density f_{α_0,β_0}. From the generated data, using Part 1.2, get three estimates of $q_{0.95}$.

Hint : The following result is useful to generate an *iid* sample from any density function : let $U = F(T)$, where F is the cumulative distribution function of a continuous random variable T. Hence, we can prove that $U \sim Unif[0,1]$.

(b) Repeat this data generating process $N = 100$ times (with the same sample size $n = 20$ and the same (α_0, β_0)). Hence, you obtain a sample of size N of each estimator of $q_{0.95}$. Make a histogram and a boxplot of these three samples. What can you conclude ?

(c) Use the samples obtained in (b) to estimate the bias, the variance and the mean squared error of $\widehat{q}_S$, $\widehat{q}_M$ and $\widehat{q}_L$. What can you conclude ?

(d) Repeat the calculations in (c) for $n = 20, 50, 100, 250, 400$. Compare the biases, the variances and the mean squared errors of $\widehat{q}_S$, $\widehat{q}_M$ and $\widehat{q}_L$ graphically (make a separate plot for each quantity as a function of n). What can you conclude about $\widehat{q}_S$, $\widehat{q}_M$ and $\widehat{q}_L$? Which estimator is the best ? Justify your answer.

(e) Create an histogram for $\sqrt{n}(\widehat{q}_L - q_{0.95})$, for $n = 20$, $n = 100$ and $n = 400$. What can you conclude ?

NB : Parts 1.2 and 1.3 can (and should) be done (right) after TP8.

Part 2 : Regression

Context and data description

In this part a fictitious dataset on cars has been generated. You can find the dataset «dataproject.txt» on Moodle and use the function `read.csv` to load it into R. The aim is to study the existing relationship between the logarithm of fuel efficiency measured in liters per 100 kilometers (l/100km) and horsepower (hp). The dataset consists in 100 vehicles and the variables of interest are summarized in the table given below.

Name of the variable	Description
Y	The logarithm of fuel consumption in l/100km
X	Horsepower

Questions

(a) Fit a linear regression model where Y is the response variable and X the explanatory variable. Represent the estimation results of the slope and intercept in a table and give an interpretation.

(b) 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 ?

(c) Let us assume that the linear model is valid. In 2015, Christian von Koenigsegg designed a car named Koenigsegg One :1. This name has been chosen because the car has approximately a ratio of 1 horsepower per kilogram. Given that this car has 1000 horsepower, predict the expected value of the logarithm of fuel consumption. In addition, give a 95% prediction interval for this value. Interpret your results.

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