DS4043 - Introduction to Statistical Computing - Final Project 2022

Due on May 12, 2022 at 11:59 pm

Overview

The purpose of this project is to train students to perform the tasks that require the following skills.

- Reading and researching
- Understanding and application of distributions
- Statistical analysis/programming/simulation in R
- Teamwork
- Report writing

Instruction

- Each group will be formed by 2-3 students, should indicate the group members by April 25.
- The choice of datasets is available from R website, i.e.
  https://stat.ethz.ch/Rmanual/R-devel/library/datasets/html/00Index.html.
  https://www.kaggle.com/datasets
  https://archive.ics.uci.edu/ml/datasets.php
  (other reasonable datasets are also acceptable)
- Each group needs to complete a written report by week 14 Thursday, May 12. **Each group only needs to submit one copy of R markdown file and the generated pdf.**
- The pdf report (excluding appendix) should not include more than 10 pages.

Project Report

Your reports should at least include the following parts.

- Introduction: the background of the project
- Objective: your aim and interest of the project
- Data description: data source, description and the variable/s you are interested
- Data analysis:
  - Display the distribution for the variable/s of your chosen dataset by plotting the histogram and Kernel density estimates
  - Calculate and explain the estimated mean and variance under both the sample and ML methods for the variable/s of your chosen dataset
  - Visualization of the dataset
  - You can include other data analysis methods.
- Method:
  - Explain the Kernel density estimation method
  - Explain the maximum likelihood (ML) estimation and sample (i.e. the sample mean and variance estimates) methods, for the mean and variance of the variable/s chosen, based on distribution assumption/s

- Modeling (if you have):
  - Use jackknife to help you select models (See book page 208; Example 7.17 (Model selection))
  - Use bootstrap to estimate the bias and standard error of your estimators
- Simulation:
  - Generate the Monte Carlo (MC) sample/s using the distribution assumption (based on the data analysis results) of the variables you selected
  - Estimate the mean and variance using both the sample and ML methods for the MC sample/s
  - Compare the sample and ML estimates of mean and variance using mean square errors, at both small (e.g., n=50) and large samples (e.g., n=1000)
- Conclusion: give summary and thoughts you would like to further investigate.
- Appendix: If you have some formulas, derivations and computer, codes would like to be included in the report.
- Contribution: Including group leader and each group members' contribution descriptions.
- Reference