SS4864A/SS9864A

Assignment #4 due to November 19, 2021

Instructions

Assignments must be prepared by using RStudio Markdown/Notebook with appropriately annotated with comments, plots, and explanations; notice that neatness counts. Submit your pdf output as well as Rmd (R Markdown/Notebook) files to owl before deadline.

1. In this question you will study the Least Absolute Deviation (LAD) regression (for simple linear regression) with comparison to Least Squares (LS) regression. LAD is also known as Least Absolute Errors (LAE), Least Absolute Value (LAV). Since it is difficult to do inference for LAD regression, bootstrap method will be used to estimate standard errors of LAD regression parameters. Carry out the following small simulation study.

   (a) DGP: Write a function with proper arguments to generate data sets according to

      i. x is a random vector (length $n = 31$) generated as

      ```
      set.seed(4864)
      n = 31
      x = sample(0:15, size = n, replace = TRUE)
      set.seed(as.POSIXct(Sys.time()))
      ```

      ii. $a = 3$ and $b = 0.3$.

      iii. $\varepsilon_i \sim 95\% * N(0,1) + 5\% * N(0, k^2), i = 1, 2, \ldots, n$. You must standardize it so that its variance is 1.

      iv. $y = a + b * x + \varepsilon$.

   (b) Write a function to compute LAD estimations of a and b with optimization function **nlminb**. Detail steps list below:

      i. Your function argument lists must include the pair observations x, y as either a data frame or two vectors.

      ii. In your function body, compute the LAD as $\min \sum_{i=1}^{n} |y_i - (a + b * x_i)|$.

      iii. For the initial values of a and b you need to compute the LS estimations of a and b by using **lm** function and use them as the initial values or the start point.

      iv. Use **nlminb** to find a and b which minimize $\sum_{i=1}^{n} |y_i - (a + b * x_i)|$.

      v. Return the estimated values of a and b.

   (c) Use the function in (a) to generate one set of x, y with $k = 1$. Use the function in (b) to get estimations of a, b. Use **lm** to compute LS estimations of a, b. Compare and comment.

   (d) Redo (c) with $k = 8$ (ε is called contaminated normal errors).

   (e) There is no formula available to compute standard errors for LAD estimations of a, b. Here you will use bootstrap to obtain their standard errors. In this assignment, you will resample the n pairs $(x_i, y_i), i = 1, 2, \ldots, n$. Here are procedures you should follow:

      i. Write a function with arguments R=1000 (as the number of bootstraps) and pairs of observations x, y with the same format used in (b).

ii. In the function body, resample x, y to generate a bootstrap pair sample $x*, y*$ and use (b) to get corresponding LAD bootstrap estimations of a, b.

iii. Run R times to get a vectors of those a, b estimators.

iv. Compute the standard errors of a and b respectively.

v. Return those two standard errors as well as the LAD estimators based on the original pair x, y.

vi. Test your function with the pairs generated in (c) and (d). Report your findings.

(f) Repeat the above $K = 1000$ times for standard ($k = 1$) and contaminated normal errors ($k = 8$), i.e., generate K independent samples in (c) as well as in (d). Examine the respective LAD estimates and bootstrap estimates of standard errors and comment. How do LAD estimates compare to LS estimates? How do standard errors of LAD estimates compare to those of LS estimates? To be more specific, you need at least to examine

i. Biases and standard errors of LAD and LS estimates from K estimations

ii. Mean and standard errors of K standard errors of LAD and LS estimates.

iii. Does the means in **ii** match those standard errors in **i** respectively?

iv. Use density plots to compare those estimates.

(g) LAD estimation is relatively slow so the above simulation 1000x1000 ($K * R$) may take some time to finish, depending on the type of CPU you are using. Please use **system.time** to record the total time to finish a set of simulation in (f). Then use R's **parallel** package to parallel the codes (for $K = 1000$ simulation part and hence needing to modify the function in (e)) and record the total time to finish a set of simulation. Make sure you enable parallel RNG before simulation. Once you find that the parallel procedure will speed up the simulation, rerun (f) with $K = 2000$ and $R = 2000$. Comment the statistical differences between the simulations 1000x1000 and 2000x2000.

2. Redo part of Question 3 from Assignment 2 in parallel way. Reuse or modify functions already implemented like **statistic.star** and **test.ci**. It is a parallel Monte Carlo study of two bootstrap based confidence intervals in term of bias coverage. Inputs are **x, statistic, theta0, R**=50000. The confidence level is fixed at 95% though you can change it to a general level $1 - \alpha$ and put it as one of inputs.

(a) Implement a function **test.ci** with arguments **x, statistic, theta0, R**=50000, where **theta0** is the true parameter value. In the function body, first use the function **statistic.star** (from Q3(a) in Assignment 2) to generate a output (a vector with those **statistic**(x^*)). Then use this output to construct three confidence intervals:

i. Construct a c.i. based on normal approximation: statistic(x) $\pm 1.96 * sd$(bootstrap output). Calculate the bias as: the center of c.i. - **theta0**.

ii. Construct a c.i. based on basic percentile method given in class. Calculate the bias as: the center of c.i. - **theta0**.

iii. Construct a c.i. pretending that the distribution of statistic(x^*) is the same as that of statistic(x) and use the lower and upper of percentiles of statistic(x^*) to construct the c.i. Calculate the bias as: the center of c.i. - **theta0**.

Finally your function should return a vector of the above 3 values.

(b) Carry out two parallel Monte Carlo simulations with size K=50000. The population is the Exponential distribution with rate 1/4 and the statistic is the sample mean.. The sample size for the first Monte Carlo size is 15 and the second is 50. Conclude which c.i. gives the less bias coverage? Comments your findings.

3. Not yet.