Home Assignment: Statistical Computing with R (ST3004)

Yujiao Li, Murshid Saqlain & Diala Jomaa

Deadline April 7, 2019

Instructions

This home assignment contains 3 questions. Your answers should be complete, clearly written and easy to go through. All the necessary calculations and mathematical derivations should be presented in details. Necessary R source codes related to your solutions should be clearly documented and sufficiently commented. Supply the R source code as a supplementary document (in *.R or *.txt format). The main report, containing your solutions, should contain a cover page which shows your name, e-mail address and registration/person number. Your solution should be done individually by yourself without requiring any help from any other person or resources. In case of using others' published work, you should provide proper reference to the original source.

Your final solution should be submitted by April 7, 2019, 12:00 CET.

Good Luck!

Re-sampling methods

1. Suppose you have a pack of poker with 52 cards. Among the 52 cards, there are 4 suits (spades, clubs, diamonds and hearts) and 13 values for each suit. One can draw a sample of 5 cards from the pack each time. We wonder the probability of getting at most two different suits in one drawing. (20p)

   (a) Write a function in R. The input is the times of draws (n) and the output is the probability.

   (b) Plot how the probability varies as n goes from 1 to 500.

   (c) Repeat 5, 50 and 500 times for n=500 and then separately calculate their mean value.

Regression

2. For this question you need the 2008 Airline data you can find here `http://stat-computing.org/dataexpo/2009/the-data.html` if you not already have downloaded it. (40p)

   (a) Import the data as an *ff* data frame (for this you will need the packages *ff* and *ffbase*).

   (b) Take a subset of the data for the month December. Henceforth you will use this subset for the analysis. This should not be so big that you cannot fit it in the memory, so you can work with the traditional functions in R (e.g. lm()).

   (c) If you have the subset in a *ff* data.frame make it into a traditional R data.frame and show that your data object actually is a R data.frame.

   (d) Obtain the sample size of this data.

   (e) Run a regression with the following formula:

   ```
   ActualElapsedTime ~ DayOfWeek+UniqueCarrier+Distance
   ```

   (f) What is the predicted average travel time for Bob who is going to travel 1000 miles, next Thursday with Pinnacle Airlines Inc. (9E)?

(g) What is the predicted average travel time for Susanna who is going to make the same trip with the same carrier as Bob but on Saturday?

(h) Obtain heteroskedatic-robust t-tests for the regression.

(i) Does the predicted travel time (statistically) significantly differ between Bob's trip on a Thursday and Susanna's trip on a Saturday? Explain shortly how you determine this.

Regression and Big Data

3. Continue to work with the Airline data but now the complete data set for both 2007 and 2008 (If it is very tedious to work with the complete data, a suggestion is to take a subset of it and construct appropriate R-code and finally run it on the whole sample during night.) (40p)

   (a) Obtain the sample size of this data.

   (b) Run two separate regressions with the following formula for the two years 2007 and 2008:

   ```
   ArrDelay~ Origin + Month + Distance
   ```

   using a function presented during the course for handling big data.

   (c) What was the predicted delay for Sue who was going 10000 miles on Christmas Eve in 2008 from Newark Airport (EWR)? Given the results from the regression in 3b.

   (d) With the *pred_airports* function, you can find in the script **Predict_airports_exam.R**(attached), you can obtain the predicted values for each airport and month holding the distance fixed at the sample average. Try to obtain the predicted value for each airport, month and year. Do it separately for each year but put the results in one data.frame. Make sure you have a variable called "year" that gives the year of each predicted value.

   Remark: This is not the predict function we used in the exercise set for big data and you cannot use that script. You need to read the instruction in the script *Predict_airports_exam.R*, to see how you are going to use the function. If you get it to work you should obtain a data frame including airport abbreviations, month and predicted values (pred). The predict function only

works with a Chunk_lm object not with an object from biglm(). It will be possible to adjust it to work with biglm() objects too but then you have to modify the function yourself.