Nathalie Vialaneix
Année 2018/2019

M1 in Economics and Economics and Statistics
Applied multivariate Analysis - Big data analytics
Worksheet 1 - Bootstrap

This worksheet illustrates the use of nonparametric bootstrap to estimate confidence intervals and variance of estimators. The simulations use R. An introduction to R, to basic use of packages and of the RStudio environment and useful references to obtain help can be found at http://www.nathalievialaneix.eu/teaching/tide. In particular, any time you do not understand how to use a function, you must use the help:

```r
help(rgamma)
```

to have the help page for the function `rgamma` and carefully read it.

Exercice 1 My first bootstrap estimation

This exercise's aim is to program a boostratp estimate of the confidence interval of the mean.

1. Using the function `rgamma`, generate a random sample from the Γ distribution with parameters $k = 2$ and $\theta = 1$, having size $n = 15$. We recall that the density of $\Gamma(k, \theta)$ is:

$$f_{k,\theta}(x) = x^{k-1} \frac{\exp(-x/\theta)}{\Gamma(k)\,\theta^k} \mathbb{I}_{\{x>0\}}$$

   with Γ the Γ function $\Gamma(k) = (k-1)!$, $\forall\, k \in \mathbb{N}^*$. What is the mean of the $\Gamma(k, \theta)$ distribution? What is the empirical mean of your sample?

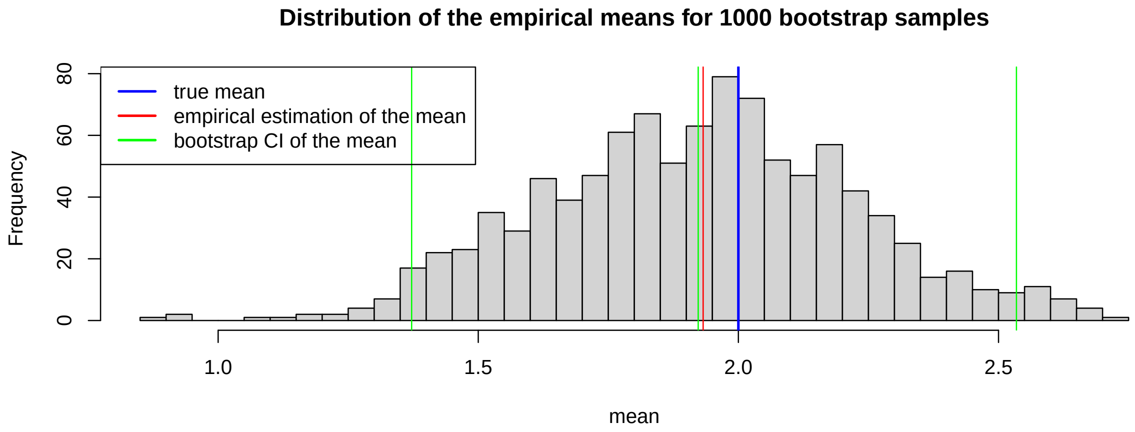
2. Using the function `sample`, create a function that:

   - takes a given sample (vector) for input;
   - generates a bootstrap sample from the previous sample;
   - outputs its (empirical) mean.

   Test this function to obtain the mean from a bootstrap sample by using the sample generated in the previous question.

3. Use this function in a `for` loop to obtain 1000 values of mean from bootstrap samples. What is the bootstrap estimate of the mean? The bootstrap estimate of the 2.5% and 97.5% of the quantiles of this distribution?

4. Draw the distribution of these means and identify the true mean, the empirical estimateof the mean, the bootstrap estimate of the mean and the bootstrap 95% confidence interval of the mean.



Exercice 2 Bootstrap using the package boot

In this exercise, the package **boot** is used to calculate a bootstrap estimate of the variance of the empirical estimate. The computational time used by the package **boot** is compared to the use of a `for` loop.

1. Using the function `rbeta`, generate a random sample from the Beta(α,β) distribution with parameter $\alpha = 1$ and $\beta = 2$[1], having size $n = 20$. What is the expected mean of the Beta$(1, 2)$ distribution? What is the empirical estimate, $\bar{x}$, of this mean for the generated sample?

2. What is the variance of the estimator $\bar{x}$?

3. Using the sample generated in the previous question and a `for` loop, find a bootstrap estimate for the variance of $\bar{x}$ with $B = 1000$ bootstrap samples. Use the function `system.time` to obtain the computational time required to compute the 10000 bootstrap estimates of the mean.

4. Load the package **boot** with the command line:

   ```
   library(boot)
   ```

   The function `boot` requires an argument `statistic` which is a function having two arguments, the original sample and a vector of indexes for a subsample which calculates the empirical estimate (here the mean) from the bootstrap sample. Use

   ```
   ?boot
   ```

   to see examples of such functions and create a function

   ```
   boot.mean <- function(a.sample, sub.indexes) {
     ...
   }
   ```

   which calcultes the empirical mean from the bootstrap sample with observations `sub.indexes`.

5. Use the `boot` function to generate 1000 estimates of the mean from bootstrap samples. Store this results in a R object named `res.boot`. Print `res.boot` and use `names(res.boot)` to see which information is included in this object (also use the help page of the function). How to obtain a bootstrap estimate of the variance of $\bar{x}$ from `res.boot`?

6. Compare the computational time required by the `boot` function to the computational time required by the method used in question 3 for 10000 bootstrap estimates.

[1] We recall that the density of the Beta(α,β) distribution is $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}_{[0,1]}(x)$ whose mean is equal to $\frac{\alpha}{\alpha+\beta}$ and whose variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

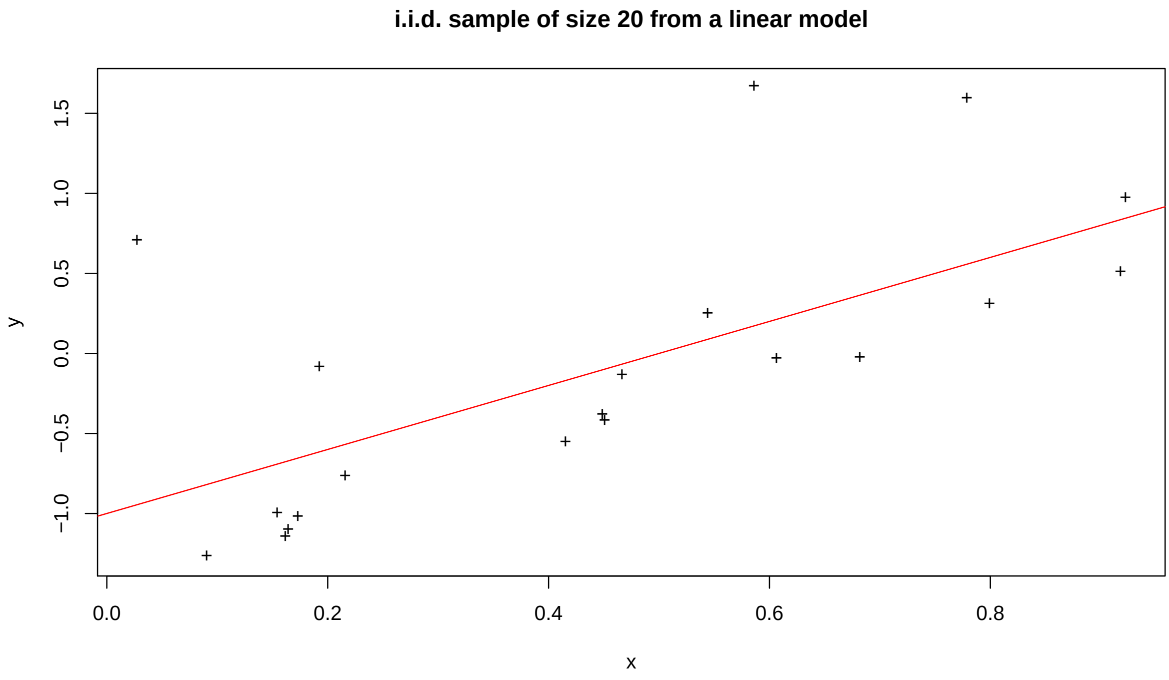
Exercice 3 Bootstrap in linear models

This exercise illustrates how the bootstrap can be used in the context of a linear model to obtain confidence intervals for the parameters and confidence bounds for the prediction.

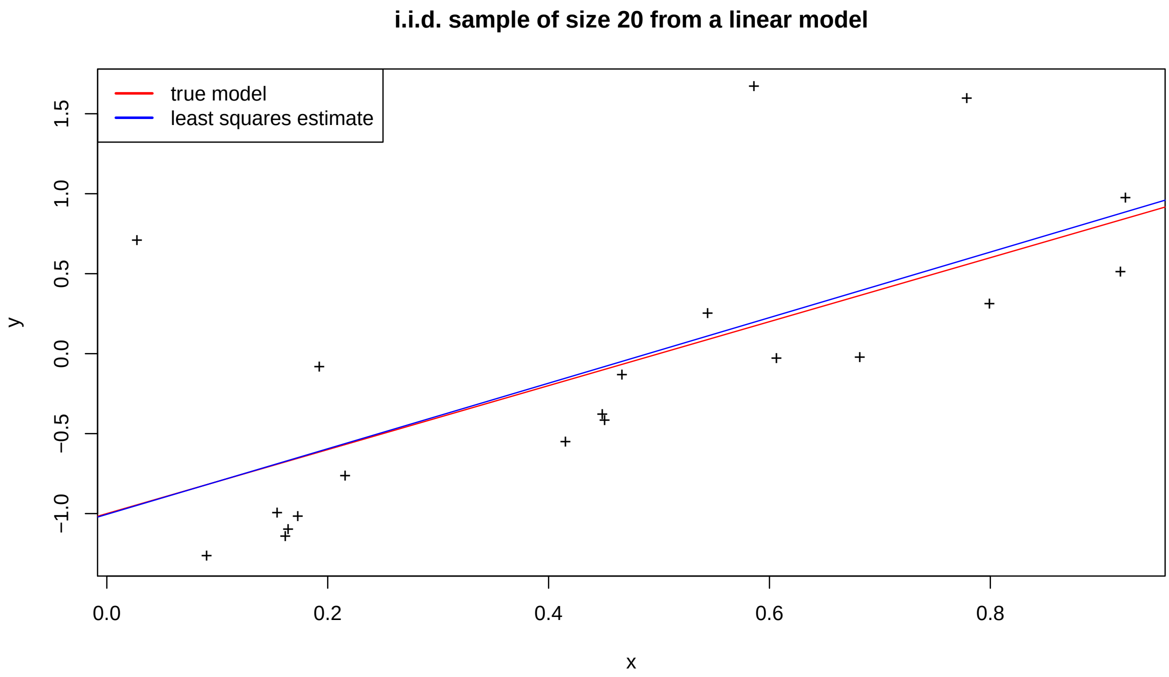
1. Generate 20 i.i.d. observations from $X \sim \mathcal{U}[0,1]$ where $\mathcal{U}[0,1]$ is the uniform distribution in $[0,1]$. Then generate from this sample, 20 observations from Y, which Y fitting the model

$$Y = 2X - 1 + 0.5(\epsilon - 1),$$

where $\epsilon \sim \chi^2(1)$ and ϵ and X are independant. Note that, for this model, the law of the least squares estimates of the parameter ($\alpha = 2$ and $\beta = -1$) of the model is not the usual normal distribution. Make a graphic with the i.i.d. observations of the couple (X, Y) and add the true linear model on it (*i.e.*, the line $y = 2x - 1$).

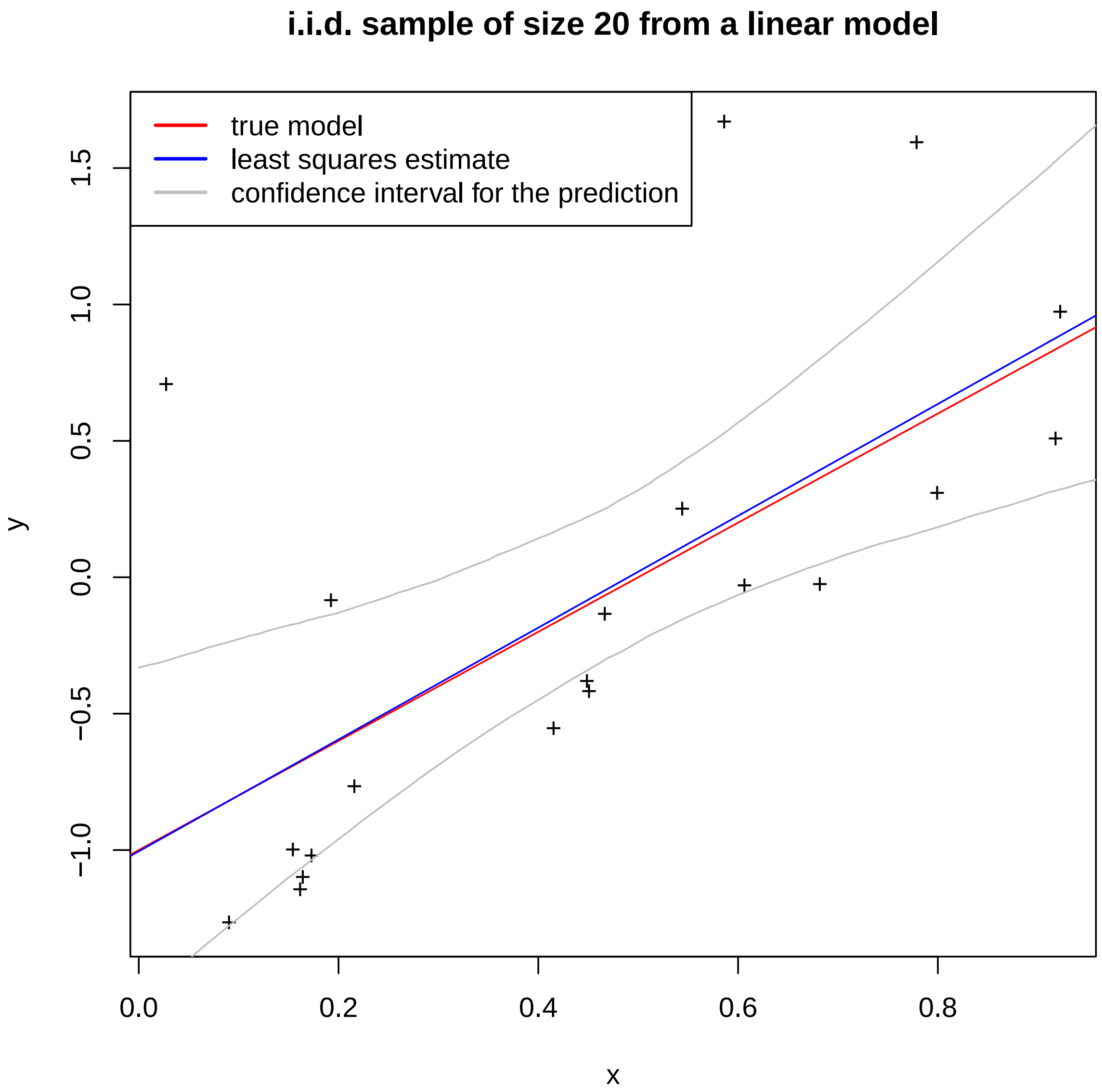


2. Use the function `lm` to find the least squares estimates of the parameters, $\hat{\alpha}$ and $\hat{\beta}$ of the model. Add the estimated linear model to the figure of the previous questions.



3. Use the function `boot` to find a 95% confidence interval for $\hat{\alpha}$ and $\hat{\beta}$ with 5000 bootstrap samples.

4. Use the function `boot` to obtain 95% confidence intervals for the estimations of the predictions associated to $x = 0$, 0.01, 0.02, ..., 1. Add the bounds of the predictions to the figure obtained in question 2.



i.i.d. sample of size 20 from a linear model
true model
least squares estimate
confidence interval for the prediction
y
1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
x