Bootstrap

Jawad Boulahfa

06/05/2020

Deuxième partie: bootstrap

```
rm(list = ls())
```

Installation du package test

```
#devtools::install_github("Jawad-Boulahfa/test")
```

Chargement des packages

```
# Pour pcls2 et les fonctions de simulation
library(test, quietly = TRUE)

# Pour le calcul parallèle
library(foreach, quietly = TRUE)
library(iterators, quietly = TRUE)
library(parallel, quietly = TRUE)
library(doParallel, quietly = TRUE)

# Pour ggplot2 et la manipulation des dataframes
library(tidyverse, quietly = TRUE)

# Pour tracer plusieurs graphiques en même temps
library(gridExtra, quietly = TRUE)

# Pour construire des heatmap
library(reshape2, quietly = TRUE)

# Pour construire des heatmap
library(hrbrthemes)
library(plotly)
library(webshot)
```

Initialisation

```
n <- 1000 # Nombre de lignes
sigma <- 1 # Ecart-type du bruit gaussien
beta <- c(0, 1) # Choix du beta

prop <- 0.7 # Proportion de données prises pour rééchantillonner
```

```r
replace <- TRUE # On autorise les répétitions dans le rééchantillonnage

B <- 200 # Nombre de fois où on répète le réechantillonnage (pour le bootstrap)

alpha <- 0.05 # Pour les intervalles de confiance à 95%

nb_classes <- 50 # Nombre de classes pour les histogrammes

# Nombre de fois où on refait un bootstrap
# pour construire un nouvel intervalle de confiance
# Autrement dit, on construira 100 IC ici (via 100 bootstrap)
nb_repetitions <- 100
```

Premier essai

On calcule $\hat{\beta}_{nnls}$ et $\hat{\beta}_{lm}$ pour chaque jeu de données rééchantillonné (*nombre de rééchantillonnages* = 200, $n = 1000$, $\sigma = 1, \beta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$).

```r
essai <- test::beta_list(n = n, sigma = sigma,
                         beta = beta, prop = prop,
                         replace = replace, B = B)
```

On affiche un aperçu des résultats obtenus.

```r
print(essai, max = 10)
```

```
## $beta_nnls
##   beta_nnls_1 beta_nnls_2
## 1           0   0.8040702
## 2           0   0.8805094
## 3           0   1.0223057
## 4           0   0.9781424
## 5           0   0.8870369
##  [ reached 'max' / getOption("max.print") -- omitted 195 rows ]
##
## $beta_lm
##     beta_lm_1 beta_lm_2
## 1 -0.16641628 0.9237064
## 2 -0.26430952 1.0788666
## 3 -0.15783630 1.1413139
## 4 -0.14436609 1.0885021
## 5 -0.09887137 0.9638879
##  [ reached 'max' / getOption("max.print") -- omitted 195 rows ]
##
## $comparison_1
##   beta_1 model
## 1      0  nnls
## 2      0  nnls
## 3      0  nnls
## 4      0  nnls
## 5      0  nnls
##  [ reached 'max' / getOption("max.print") -- omitted 395 rows ]
##
## $comparison_2
```

```
##      beta_2 model
## 1 0.8040702  nnls
## 2 0.8805094  nnls
## 3 1.0223057  nnls
## 4 0.9781424  nnls
## 5 0.8870369  nnls
##  [ reached 'max' / getOption("max.print") -- omitted 395 rows ]
##
## $biais_df
##               biais
## nnls_1  0.003480749
## lm_1   -0.146425850
## nnls_2 -0.107327993
## lm_2    0.007457054
##
## $var_df
##            variance
## nnls_1 0.0002285577
## lm_1   0.0097375385
## nnls_2 0.0051276805
## lm_2   0.0106533509
##
## $MSE_df
##                 MSE
## nnls_1 0.0002406734
## lm_1   0.0311780680
## nnls_2 0.0166469787
## lm_2   0.0107089585
##
## $IC_df
##           IC_nnls_1    IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.3322365 0.7411099 0.8063268
## Borne sup 0.0537995  0.0537995 1.0224263 1.1996976
```

La variable essai contient:

- un dataframe contenant les valeurs de $\hat{\beta}_{nnls_1}$ et $\hat{\beta}_{nnls_2}$
- un dataframe contenant $\hat{\beta}_{lm_1}$ et $\hat{\beta}_{lm_2}$
- un dataframe contenant les valeurs de $\hat{\beta}_{nnls_1}$ et $\hat{\beta}_{lm_1}$ regroupées sur une colonne (et qu'on peut distinguer à l'aide de la colonne "model") un dataframe contenant les valeurs de $\hat{\beta}_{nnls_2}$ et $\hat{\beta}_{lm_2}$ regroupées sur une colonne (et qu'on peut distinguer à l'aide de la colonne "model")
- un dataframe contenant le biais de chaque composante des deux estimateurs.
- un dataframe contenant la variance de chaque composante des deux estimateurs.
- un dataframe contenant l'erreur quadratique moyenne de chaque composante des deux estimateurs
- un dataframe contenant les intervalles de confiance au niveau de confiance 0.95% de chaque composante des deux estimateurs.

Histogrammes et indicateurs

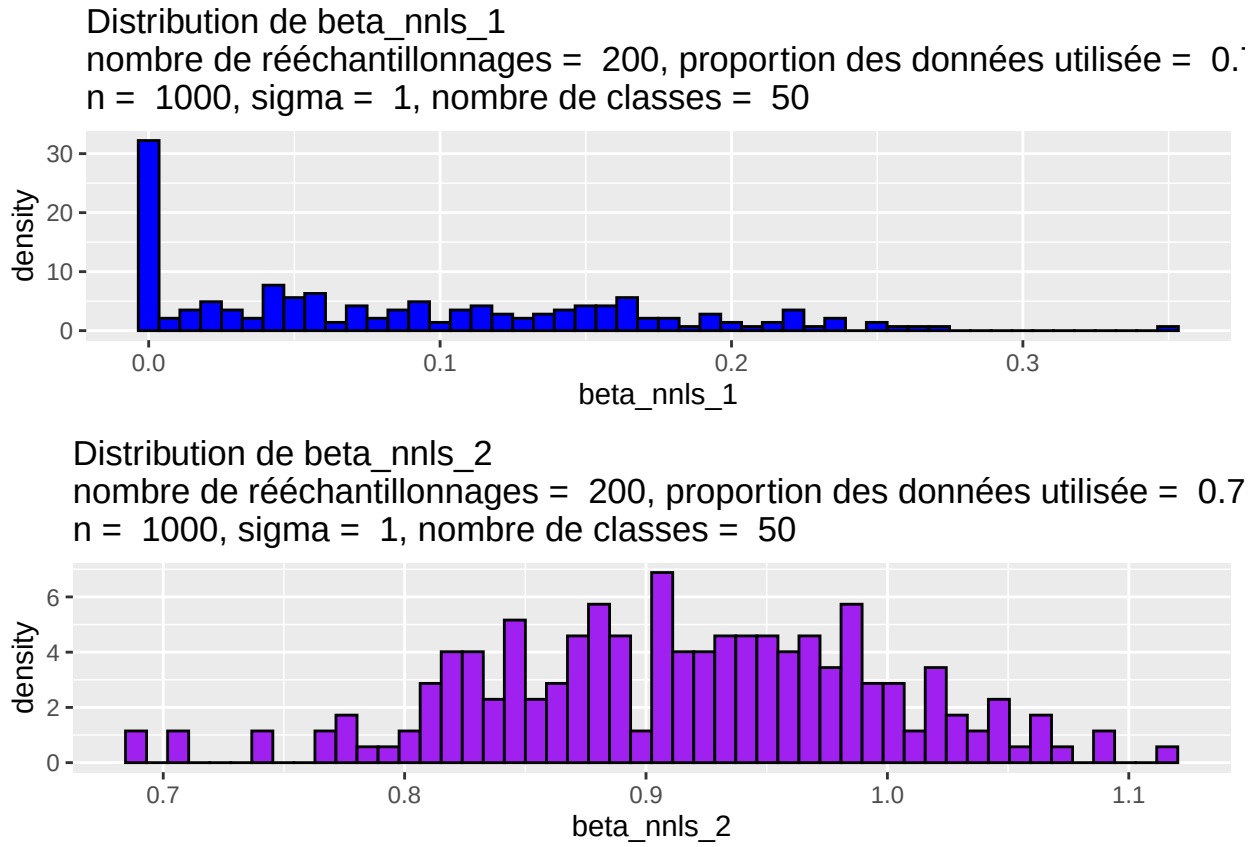
```
distribution <- test::distribution_beta(
  n = n, sigma = sigma, beta = beta,
```

```
  prop = prop, replace = replace, B = B)
```

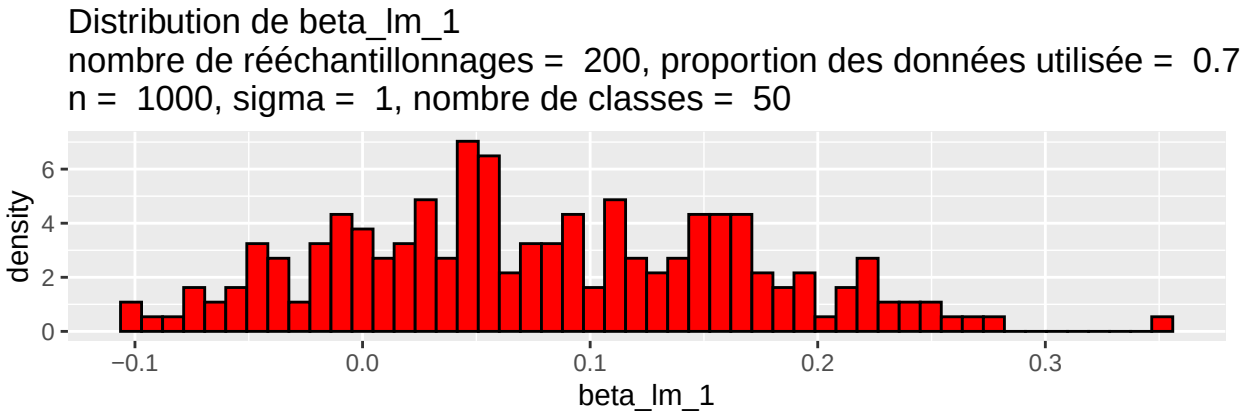
Affichage des 4 histogrammes séparément

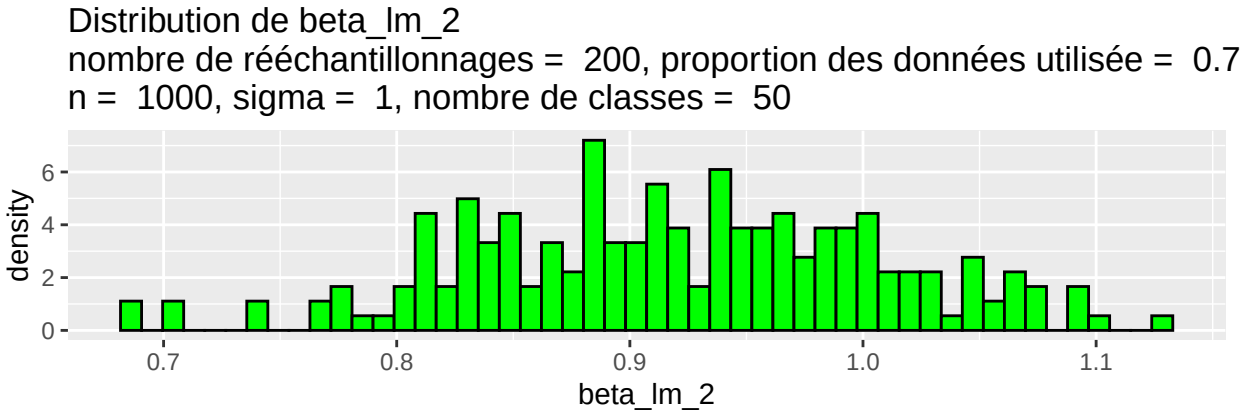
On affiche la distribution de chacune des composantes de $\hat{\beta}_{nnls}$ et $\hat{\beta}_{lm}$ ($nombre\ de\ rééchantillonnages = 200,\ n = 1000,\ \sigma = 1$).

```
grid.arrange(distribution$beta_nnls_1_hist,
             distribution$beta_nnls_2_hist,
             nrow = 2, ncol = 1) # OK
```



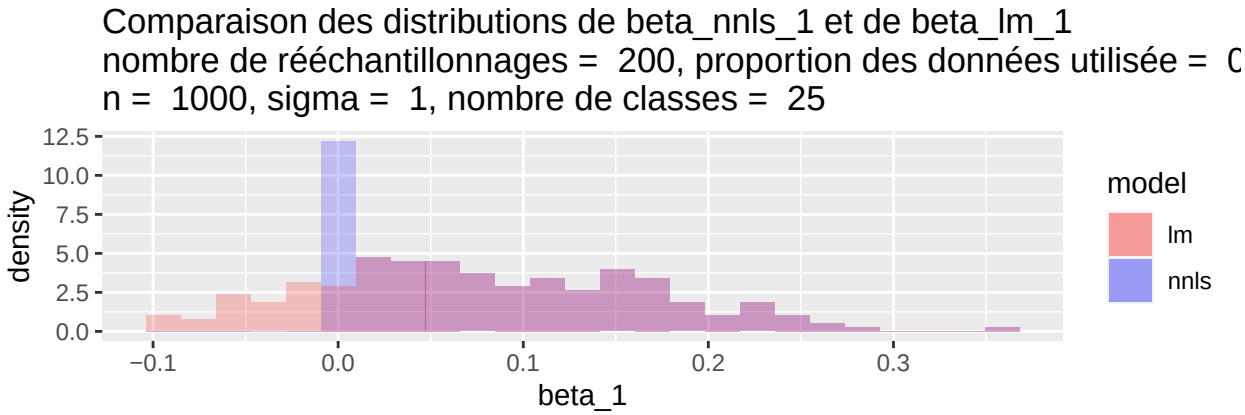

```
grid.arrange(distribution$beta_lm_1_hist,
             distribution$beta_lm_2_hist,
             nrow = 2, ncol = 1) # OK
```

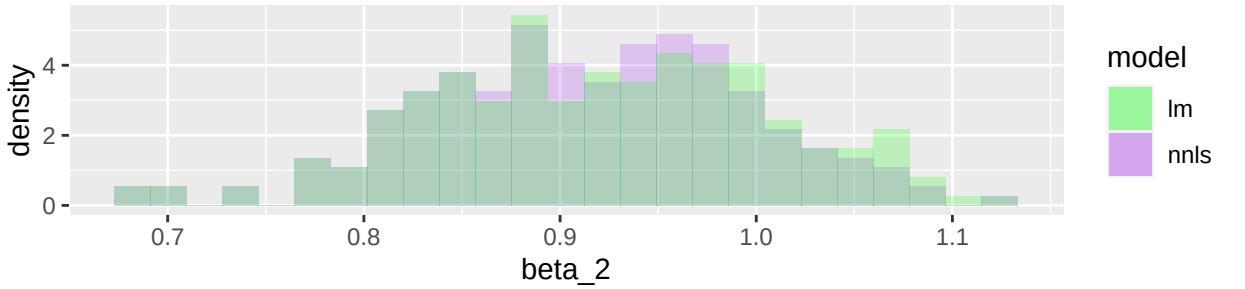




Comparaisons entre les modèles nnls et lm

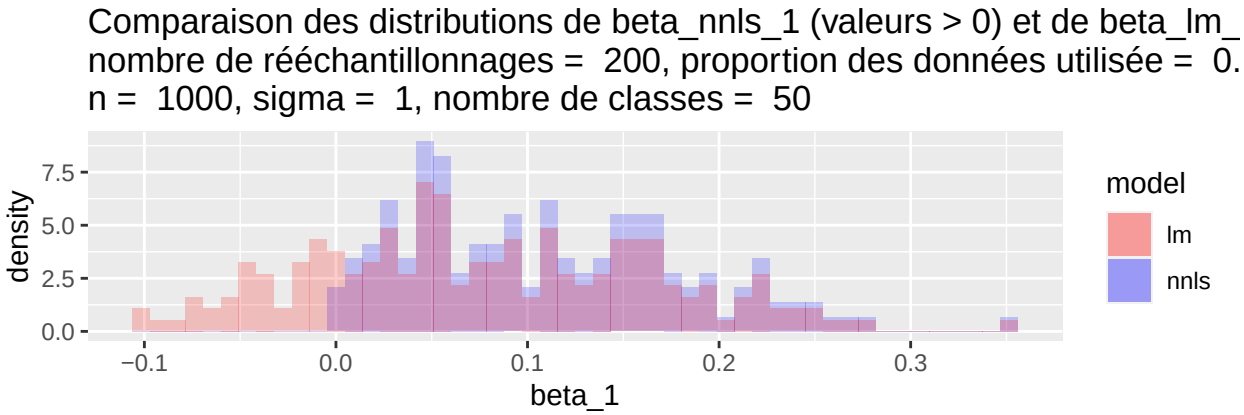
```
grid.arrange(distribution$comparison_1_hist,
             distribution$comparison_2_hist,
             nrow = 2, ncol = 1)
```

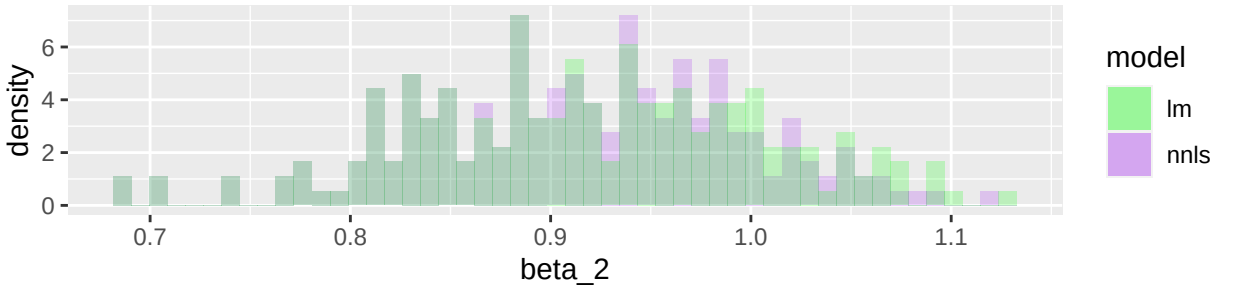




Comparaisons sans le pic en 0 pour nnls

```
grid.arrange(distribution$comparison_1_hist_without_0,
             distribution$comparison_2_hist_without_0,
             nrow = 2, ncol = 1)
```





Biais, variance, erreur quadratique moyenne

On affiche le biais, la variance, et l'erreur quadratique moyenne de chaque composante de chacun des estimateurs.

```
# Biais
print(distribution$biais_df)
```

```
##              biais
## nnls_1  0.08586761
## lm_1    0.07776402
## nnls_2 -0.08548317
## lm_2   -0.07953768
```

```
# Variance
print(distribution$var_df)
```

```
##           variance
## nnls_1 0.005993451
## lm_1   0.007806929
## nnls_2 0.006708963
## lm_2   0.007641191
```

```
# MSE
print(distribution$MSE_df)
```

```
##               MSE
## nnls_1 0.01336670
```

```
## lm_1   0.01385417
## nnls_2 0.01401634
## lm_2   0.01396743
```

On peut rassembler tous ces résultats dans un seul dataframe pour plus de lisibilité.

```
biais_var_MSE_df <- cbind(distribution$biais_df,
                          distribution$var_df,
                          distribution$MSE_df)
```

```
print(biais_var_MSE_df)
```

```
##              biais    variance        MSE
## nnls_1  0.08586761 0.005993451 0.01336670
## lm_1    0.07776402 0.007806929 0.01385417
## nnls_2 -0.08548317 0.006708963 0.01401634
## lm_2   -0.07953768 0.007641191 0.01396743
```

Intervalles de confiance au niveau de confiance 0.95%

Intervalles de confiance pour $\beta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

```
print(distribution$IC_df)
```

```
##           IC_nnls_1     IC_lm_1 IC_nnls_2  IC_lm_2
## Borne inf 0.0000000 -0.07717584  0.743472 0.743472
## Borne sup 0.2467537  0.24675368  1.064849 1.077065
```

Rapport d'amplitude des intervalles de confiance $\beta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

```
coeff_1 <- diff(distribution$IC_df$IC_nnls_1)/
  diff(distribution$IC_df$IC_lm_1)
coeff_2 <- diff(distribution$IC_df$IC_nnls_2)/
  diff(distribution$IC_df$IC_lm_2)
```

```
print(coeff_1)
```

```
## [1] 0.7617511
```

```
print(coeff_2)
```

```
## [1] 0.9633808
```

L'amplitude de IC_nnls_1 vaut 0.7617511 fois celle de IC_lm_1. L'amplitude de IC_nnls_2 vaut 0.9633808 fois celle de IC_lm_2.

Ainsi, les intervalles de confiance obtenus pour chaque composante de $\hat{\beta}_{nnls}$ ont une plus petite amplitude que ceux obtenus pour chaque composante de $\hat{\beta}_{lm}$.

Intervalles de confiance pour $\beta \in \left\{ \begin{pmatrix} 0 \\ k \end{pmatrix} \mid k \in \{0, 1k' \mid 0 \leq k' \leq 10\} \right\}$

On construit une liste de beta.

```
nb_beta <- 11
end <- 1
```

```r
liste_beta_1 <- rep(0, nb_beta)

liste_beta_2 <- seq(from = 0, to = end, by = end/(nb_beta-1))

list_of_beta <- vector("list", length = nb_beta)

for(i in 1:nb_beta)
{
  list_of_beta[[i]] <- c(liste_beta_1[[i]], liste_beta_2[[i]])
}
```

On affiche la liste.

```r
print(list_of_beta)
```

```
## [[1]]
## [1] 0 0
##
## [[2]]
## [1] 0.0 0.1
##
## [[3]]
## [1] 0.0 0.2
##
## [[4]]
## [1] 0.0 0.3
##
## [[5]]
## [1] 0.0 0.4
##
## [[6]]
## [1] 0.0 0.5
##
## [[7]]
## [1] 0.0 0.6
##
## [[8]]
## [1] 0.0 0.7
##
## [[9]]
## [1] 0.0 0.8
##
## [[10]]
## [1] 0.0 0.9
##
## [[11]]
## [1] 0 1
```

On calcule les intervalles de confiance à 0.95%.

```r
liste_IC <- IC_beta(list_of_beta = list_of_beta,
                    n = n, sigma = sigma,
                    prop = prop,
                    replace = replace, B = B) # OK
```

On affiche les intervalles de confiance obtenus.

```r
print(liste_IC)
```

```
## $`beta = (0, 0)`
##             IC_nnls_1    IC_lm_1   IC_nnls_2    IC_lm_2
## Borne inf 0.000000000 -0.2442977 0.000000000 -0.2868992
## Borne sup 0.008449771  0.1543439 0.004630409  0.1112276
##
## $`beta = (0, 0.1)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2     IC_lm_2
## Borne inf 0.0000000 -0.1308862 0.0000000 -0.05664863
## Borne sup 0.2390158  0.2480053 0.2777288  0.34588910
##
## $`beta = (0, 0.2)`
##           IC_nnls_1    IC_lm_1  IC_nnls_2    IC_lm_2
## Borne inf  0.000000 -0.1166467 0.06892223 0.06892223
## Borne sup  0.245375  0.2453750 0.40435264 0.44962697
##
## $`beta = (0, 0.3)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.2144869 0.1427853 0.1456332
## Borne sup 0.1664901  0.1664901 0.4542898 0.5246938
##
## $`beta = (0, 0.4)`
##           IC_nnls_1     IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.09569039 0.1130300 0.1130300
## Borne sup 0.2821422  0.28214217 0.4736478 0.5201615
##
## $`beta = (0, 0.5)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.2895256 0.2364836 0.2651534
## Borne sup 0.1360445  0.1360445 0.5360163 0.6558448
##
## $`beta = (0, 0.6)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.2203378 0.4065299 0.4069175
## Borne sup 0.1698043  0.1698043 0.6963627 0.8242549
##
## $`beta = (0, 0.7)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.0000000 -0.1850410 0.3716625 0.3716625
## Borne sup 0.1820413  0.1820413 0.6662775 0.7421281
##
## $`beta = (0, 0.8)`
##           IC_nnls_1    IC_lm_1 IC_nnls_2  IC_lm_2
## Borne inf 0.0000000 -0.1591419 0.6459720 0.645972
## Borne sup 0.2541579  0.2541579 0.9499971 1.003742
##
## $`beta = (0, 0.9)`
##            IC_nnls_1     IC_lm_1 IC_nnls_2   IC_lm_2
## Borne inf 0.00000000 -0.36229806 0.8082695 0.8734623
## Borne sup 0.07186265  0.07186265 1.0641605 1.2527181
##
## $`beta = (0, 1)`
##            IC_nnls_1     IC_lm_1 IC_nnls_2   IC_lm_2
```

```
## Borne inf 0.00000000 -0.30434125 0.7797496 0.8125615
## Borne sup 0.09129141  0.09129141 1.0567774 1.2135531
```

Tests

Tests pour $\beta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

On effectue 100 bootstrap et on construit un intervalle de confiance à chaque bootstrap. Ensuite, nous effectuons deux tests. On teste $\beta_1 = 0$ contre $\beta_1 \neq 0$ pour les deux modèles en utilisant les intervalles de confiance construits pour $\hat{\beta}_{nnls_1}$ et $\hat{\beta}_{lm_1}$. On teste $\beta_2 = 0$ contre $\beta_2 \neq 0$ pour les deux modèles en utilisant les intervalles de confiance construits pour $\hat{\beta}_{nnls_2}$ et $\hat{\beta}_{lm_2}$.

```
results_test_beta <- test_beta(
  nb_repetitions = nb_repetitions,
  n = n, sigma = sigma, beta = beta,
  prop = prop, replace = replace,
  B = B, alpha = alpha,
  nb_classes = nb_classes)
```

On affiche les résultats obtenus.

```
print(results_test_beta)
```

```
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.02                 1            0.04               1
```

Les deux modèles détectent ici aussi bien la non-nullité de β_2. Néanmoins, le modèle nnls est meilleur pour détecter la nullité de β_1. On va maintenant effectuer davantage de tests en changeant les valeurs de β et commenter les résultats obtenus.

Tests pour $\beta \in \left\{ \begin{pmatrix} 0 \\ k \end{pmatrix} \mid k \in \{0, 1k' \mid 0 \leq k' \leq 10\} \right\}$

On effectue les deux tests évoqués ci-dessus, de la même manière que précédemment, mais pour chaque valeur de β dans la liste donnée en paramètre.

```
results_test_list_of_beta <- test_beta_grid(
  beta_grid = list_of_beta,
  nb_repetitions = nb_repetitions,
  n = n, sigma = sigma, prop = prop,
  replace = replace, B = B,
  alpha = alpha, nb_classes = nb_classes)
```

On affiche les résultats des tests.

```
print(results_test_list_of_beta)
```

```
## $`beta = (0, 0)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1                 0                 0            0.02            0.03
##
## $`beta = (0, 0.1)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.03               0.1            0.03            0.16
##
## $`beta = (0, 0.2)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.02              0.58            0.02             0.6
```

```
##
## $`beta = (0, 0.3)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.03              0.83            0.06            0.83
##
## $`beta = (0, 0.4)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.01              0.99            0.02            0.99
##
## $`beta = (0, 0.5)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.04                 1            0.04               1
##
## $`beta = (0, 0.6)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.01                 1            0.02               1
##
## $`beta = (0, 0.7)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.02                 1            0.02               1
##
## $`beta = (0, 0.8)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.02                 1            0.04               1
##
## $`beta = (0, 0.9)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1                 0                 1            0.01               1
##
## $`beta = (0, 1)`
##   freq_rejet_nnls_1 freq_rejet_nnls_2 freq_rejet_lm_1 freq_rejet_lm_2
## 1              0.01                 1            0.01               1
```

On remarque que la fréquence de rejet obtenue avec le modèle nnls est toujours inférieure ou égale à celle obtenue avec le modèle lm.

C'est une bonne chose lorsque $\beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, car cela montre que le modèle nnls détecte mieux la nullité des deux coefficients.

Sur les autres valeurs de β, comme on a gardé $\beta_1 = 0$, cela montre que le modèle nnls détecte également mieux la nullité d'un seul coefficient.

Néanmoins, cela signifie aussi que le modèle nnls détecte moins bien la non-nullité d'un coefficient.

En effet, ici, on souhaiterait que, lorsque $\beta_2 \neq 0$, la fréquence de rejet obtenue avec le modèle nnls soit proche voire égale à 1, et supérieure ou égale à celle obtenue avec le modèle lm.

Si la fréquence de rejet obtenue avec le modèle nnls augmente bien jusqu'à finalement atteindre 1, elle augmente moins vite que celle obtenue avec le modèle lm.

Sauvegarde des résultats

```r
save.image(file = "bootstrap_results.RData")
```

Chargement des résultats

```
rm(list = ls())
```

```
load(file = "bootstrap_results.RData")
```

```
rm(list = ls())
```