

Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #9 (bootstrap & incertitude)

Régression Linéaire & Bootstrap (1)

Dataset $\{\mathbf{z}_i = (y_i, \mathbf{x}_i)\}, i = 1, \dots, n$.

Use **paired sampling** by (repeatedly)
resampling $\{\mathbf{z}_1^\star, \dots, \mathbf{z}_n^\star\}$.

Idea :

$\{(y_i, \mathbf{x}_i)\}$ is obtained from (unknown) \mathbb{P}

Based on n observations, we observe \mathbb{P}_n

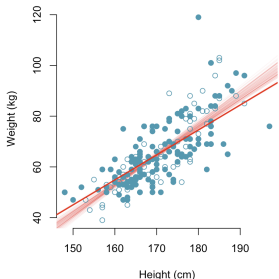
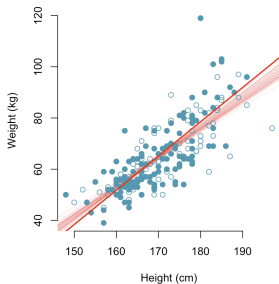
We generate other samples by resampling
from \mathbb{P}_n

1. sample $\{i_1^{(b)}, \dots, i_n^{(b)}\}$ randomly with replacement in $\{1, 2, \dots, n\}$
2. consider dataset $(\mathbf{x}_i^{(b)}, y_i^{(b)}) = (\mathbf{x}_{i^{(b)}}, y_{i^{(b)}})$'s, and fit a model
3. let $\hat{\beta}^{(b)}$ denote the estimated values, or $\hat{y}_{n+1}^{(b)}$ some prediction

Régression Linéaire & Bootstrap (1)

```
1 > BETA = matrix(NA,1000,2)
2 > for(s in 1:1000){
3   idx = sample(1:nrow(Davis),nrow(
4     Davis),replace=TRUE)
5   reg_sim = lm(weight~height, data=
6     Davis[idx,])
7   BETA[s,] = reg_sim$coefficients
8 }
9 > hist(BETA[,1])
10 > hist(BETA[,2])
```

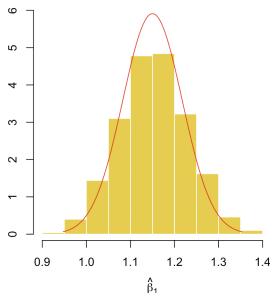
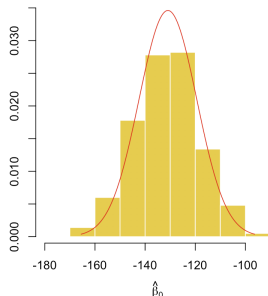
XX



Régression Linéaire & Bootstrap (1)

```
1 > BETA = matrix(NA,1000,2)
2 > for(s in 1:1000){
3   idx = sample(1:nrow(Davis),nrow(
4     Davis),replace=TRUE)
5   reg_sim = lm(weight~height, data=
6     Davis[idx,])
7   BETA[s,] = reg_sim$coefficients
8 }
9 > hist(BETA[,1])
10 > hist(BETA[,2])
```

XX



Régression Linéaire & Bootstrap (2)

As an alternative **model-based resampling**

1. sample $\widehat{\varepsilon}_1^{(b)}, \dots, \widehat{\varepsilon}_n^{(b)}$ resample from $\{\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \dots, \widehat{\varepsilon}_n\}$
2. set $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\varepsilon}_i^{(b)} = \widehat{y}_i + \widehat{\varepsilon}_i^{(b)}$
3. consider dataset $(\mathbf{x}, y^{(b)}) = (\mathbf{x}_i, y_i^{(b)})$'s and fit a model
4. let $\widehat{\beta}^{(b)}$ denote estimated values

Note in a simple regression

$$\widehat{\beta}_1^{(b)} = \frac{\sum [x_i - \bar{x}] \cdot y_i^{(b)}}{\sum [x_i - \bar{x}]^2} = \widehat{\beta}_1 + \frac{\sum [x_i - \bar{x}] \cdot \widehat{\varepsilon}_i^{(b)}}{\sum [x_i - \bar{x}]^2}$$

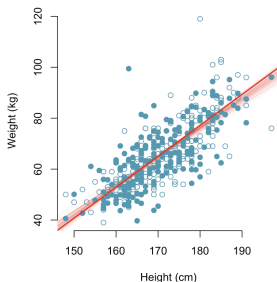
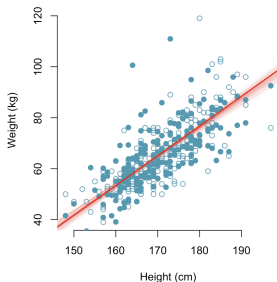
hence $\mathbb{E}[\widehat{\beta}_1^{(b)}] = \widehat{\beta}_1$, while

$$\text{Var}[\widehat{\beta}_1^{(b)}] = \frac{\sum [x_i - \bar{x}]^2 \cdot \text{Var}[\widehat{\varepsilon}_i^{(b)}]}{(\sum [x_i - \bar{x}]^2)^2} \sim \frac{\sigma^2}{\sum [x_i - \bar{x}]^2}$$

Régression Linéaire & Bootstrap (2)

```
1 > BETA = matrix(NA,1000,2)
2 > reg = lm(weight~height, data=Davis)
3 > epsilon = residuals(reg)
4 > for(s in 1:1000){
5   eps = sample(epsilon,nrow(Davis),
6               replace=TRUE)
7   Davis_s = data.frame(height =
8                       Davis$height, weight =predict(reg)
9                       +eps)
10  reg_sim = lm(weight~height, data=
11              Davis_s)
12  BETA[s,] = reg_sim$coefficients
13 }
14 > hist(BETA[,1])
15 > hist(BETA[,2])
```

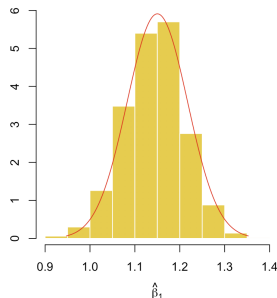
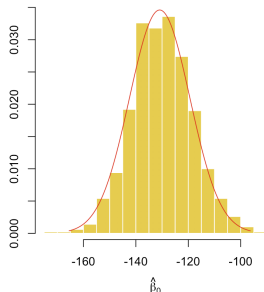
XX



Régression Linéaire & Bootstrap (2)

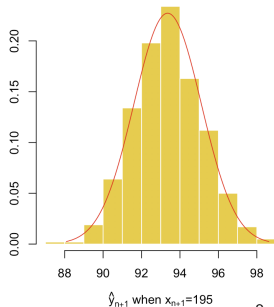
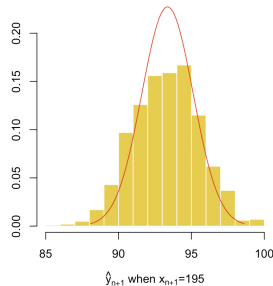
```
1 > BETA = matrix(NA,1000,2)
2 > reg = lm(weight~height, data=Davis)
3 > epsilon = residuals(reg)
4 > for(s in 1:1000){
5   eps = sample(epsilon,nrow(Davis),
6     replace=TRUE)
7   Davis_s = data.frame(height =
8     Davis$height, weight =predict(reg)
9     +eps)
10  reg_sim = lm(weight~height, data=
11    Davis_s)
12  BETA[s,] = reg_sim$coefficients
13 }
14 > hist(BETA[,1])
15 > hist(BETA[,2])
```

XX



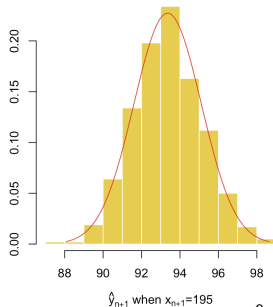
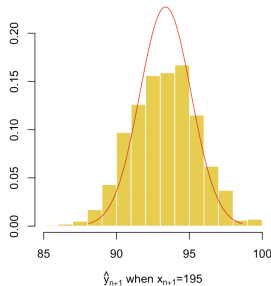
Régression Linéaire & Bootstrap (1/2)

```
1 > PRED = matrix(NA,1000,2)
2 > nwDavis = data.frame(height = 195)
3 > for(s in 1:1000){
4 +   idx = sample(1:n,n,replace=TRUE)
5 +   reg_sim = lm(weight~height, data=
6 +     Davis[idx,])
7 +   PRED[s,1] = predict(reg_sim,
8 +     newdata=nwDavis)
9 +   eps = sample(epsilon,nrow(Davis),
10 +     replace=TRUE)
11 +   Davis_s = data.frame(height =
12 +     Davis$height, weight =predict(reg)
13 +     +eps)
14 +   reg_sim = lm(weight~height, data=
15 +     Davis_s)
16 +   PRED[s,2] = predict(reg_sim,
17 +     newdata=nwDavis)
18 + }
```



Régression Linéaire & Bootstrap (1/2)

```
1 > apply(PRED,2,function(x) quantile(x
  ,.025))
2 [1] 89.04203 89.63030
3 > apply(PRED,2,function(x) quantile(x
  ,.975))
4 [1] 97.60345 97.02423
5 > predict(lm(weight~height, data=Davis
  ), newdata=nwDavis,interval="
  confidence",se.fit = TRUE)
6 $fit
7           fit           lwr           upr
8 1 93.35749 89.89571 96.81927
9
10 $se.fit
11 [1] 1.755451
```



Bootstrap heuristics

Here $\widehat{\beta} = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon = T_{\beta}(\varepsilon)$ or $\widehat{y}_{n+1} = \mathbf{x}'_{n+1} \widehat{\beta} = T_y(\varepsilon)$

Use simulations, we draw n values $\{\epsilon_1, \dots, \epsilon_n\}$ and

- ▶ $\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n T(\epsilon_i) \right] = \mathbb{E}[T(\varepsilon)]$ (unbiased)
- ▶ $\frac{1}{n} \sum_{i=1}^n T(\epsilon_i) \xrightarrow{\mathcal{L}} \mathbb{E}[T(\varepsilon)]$ as $n \rightarrow \infty$ (consistent)

Bootstrap & Tests

Consider the test of $H_0 : \beta_j = 0$,

1. compute $t_n = \frac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2}$
2. generate B bootstrap samples, under the null assumption H_0
3. for each bootstrap sample, compute $t_n^{(b)} = \frac{(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j)^2}{\widehat{\sigma}_j^{2(b)}}$
4. reject H_0 if $\frac{1}{B} \sum_{i=1}^B \mathbf{1}(t_n > t_n^{(b)}) < \alpha$.

Bootstrap & Tests

What does "generate B bootstrap samples, under the null assumption H_0 " mean ?

Example : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under H_0 , i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \dots, \widehat{\eta}_n\}$

2.2. Define $\widetilde{\eta} = \{\widetilde{\eta}_1, \dots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\frac{n}{n-1}} \widehat{\eta}$

2.3. Draw (with replacement) residuals $\widetilde{\eta}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \dots, \widetilde{\eta}_n^{(b)}\}$

2.4. Set $y_i^{(b)} = \widehat{\beta}_0 + \widetilde{\eta}_i^{(b)}$

2.5. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

3. for each bootstrap sample, compute $t_n^{(b)} = \frac{(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j)^2}{\widehat{\sigma}_j^{2(b)}}$

4. reject H_0 if $\frac{1}{B} \sum_{b=1}^B \mathbf{1}(t_n > t_n^{(b)}) < \alpha$.