Modèles Linéaires Appliqués

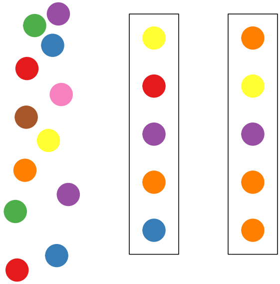
Arthur Charpentier

Automne 2020

OLS #9 (boostrap & incertitude)

Régression Linéaire & Bootstrap (1)

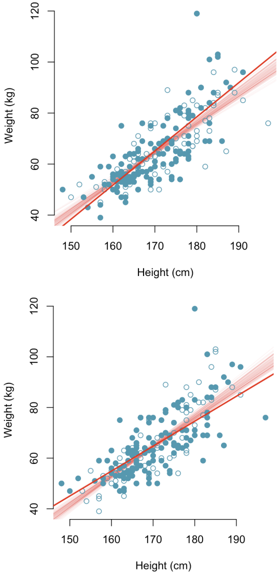
Dataset $\{\mathbf{z}_i = (y_i, \mathbf{x}_i)\}$, $i = 1, \cdots, n$.
Use paired sampling by (repeatedly) resampling $\{\mathbf{z}_1^\star, \cdots, \mathbf{z}_n^\star\}$.
Idea :
$\{(y_i, \mathbf{x}_i)\}$ is obtained from (unknown) $\mathbb{P}$
Based on n observations, we observe $\mathbb{P}_n$
We generate other samples by resampling from $\mathbb{P}_n$



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

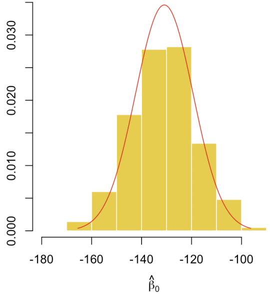
Régression Linéaire & Bootstrap (1)

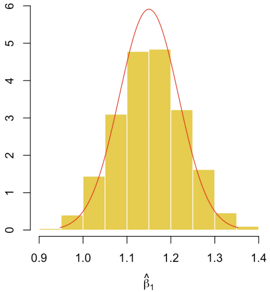
```
> BETA = matrix(NA,1000,2)
> for(s in 1:1000){
  idx = sample(1:nrow(Davis),nrow(Davis),replace=TRUE)
  reg_sim = lm(weight~height, data=Davis[idx,])
  BETA[s,] = reg_sim$coefficients
}
> hist(BETA[,1])
> hist(BETA[,2])
```



Régression Linéaire & Bootstrap (1)

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





Régression Linéaire & Bootstrap (2)

As an alternative model-based resampling

1. sample $\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)}$ resample from $\{\widehat{\varepsilon}_1,\widehat{\varepsilon}_2,\cdots,\widehat{\varepsilon}_n\}$
2. set $y_i^{(b)}=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widehat{\varepsilon}_i^{(b)}=\widehat{y}_i+\widehat{\varepsilon}_i^{(b)}$
3. consider dataset $(\mathbf{x},y^{(b)})=(\mathbf{x}_i,y_i^{(b)})$'s and fit a model
4. let $\widehat{\boldsymbol{\beta}}^{(b)}$ denote estimated values

Note in a simple regression

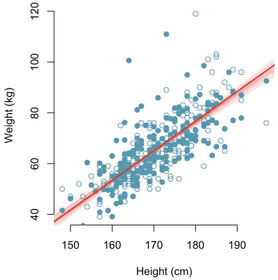
$$\widehat{\beta}_1^{(b)}=\frac{\sum[x_i-\overline{x}]\cdot y_i^{(b)}}{\sum[x_i-\overline{x}]^2}=\widehat{\beta}_1+\frac{\sum[x_i-\overline{x}]\cdot\widehat{\varepsilon}_i^{(b)}}{\sum[x_i-\overline{x}]^2}$$

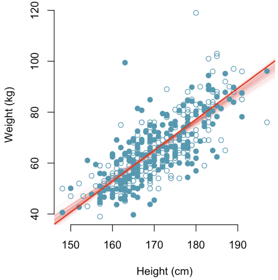
hence $\mathbb{E}[\widehat{\beta}_1^{(b)}]=\widehat{\beta}_1$, while

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

Régression Linéaire & Bootstrap (2)

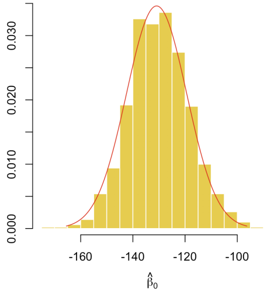
```
> BETA = matrix(NA,1000,2)
> reg = lm(weight~height, data=Davis)
> epsilon = residuals(reg)
> for(s in 1:1000){
  eps = sample(epsilon,nrow(Davis),
    replace=TRUE)
  Davis_s = data.frame(height =
    Davis$height, weight =predict(reg)
    +eps)
  reg_sim = lm(weight~height, data=
    Davis_s)
  BETA[s,] = reg_sim$coefficients
}
> hist(BETA[,1])
> hist(BETA[,2])
```

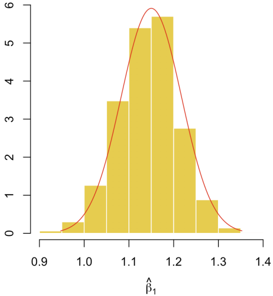




Régression Linéaire & Bootstrap (2)

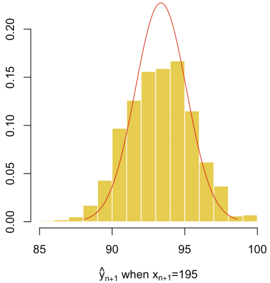
```
> BETA = matrix(NA,1000,2)
> reg = lm(weight~height, data=Davis)
> epsilon = residuals(reg)
> for(s in 1:1000){
  eps = sample(epsilon,nrow(Davis), replace=TRUE)
  Davis_s = data.frame(height = Davis$height, weight =predict(reg)+eps)
  reg_sim = lm(weight~height, data=Davis_s)
  BETA[s,] = reg_sim$coefficients
}
> hist(BETA[,1])
> hist(BETA[,2])
```

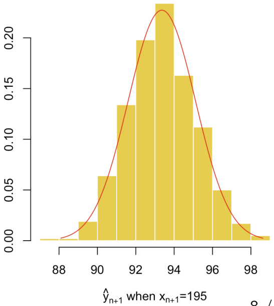




Régression Linéaire & Bootstrap (1/2)

```
> PRED = matrix(NA,1000,2)
> nwDavis = data.frame(height = 195)
> for(s in 1:1000){
+   idx = sample(1:n,n,replace=TRUE)
+   reg_sim = lm(weight~height, data=Davis[idx,])
+   PRED[s,1] = predict(reg_sim, newdata=nwDavis)
+   eps = sample(epsilon,nrow(Davis), replace=TRUE)
+   Davis_s = data.frame(height = Davis$height, weight =predict(reg)+eps)
+   reg_sim = lm(weight~height, data=Davis_s)
+   PRED[s,2] = predict(reg_sim, newdata=nwDavis)
+ }
```

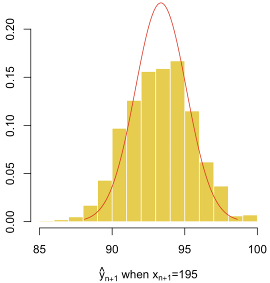


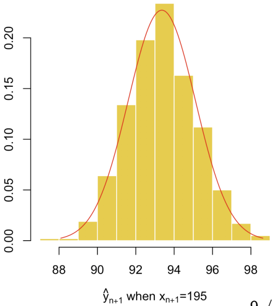


Régression Linéaire & Bootstrap (1/2)

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

$se.fit
[1] 1.755451
```





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

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

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

Bootstrap & Tests

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

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

What does "generate B boostrap 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 $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\dfrac{n}{n-1}}\widehat{\eta}$

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

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 boostrap sample, compute $t_n^{(b)} = \dfrac{(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j)^2}{\widehat{\sigma}_j^{2(b)}}$

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