Projet d'analyse de données - Baseball

Hicham, Ryad, Elyass et Hongxin

17 décembre 2022

Introduction

Le jeu de données est constitué de 322 joueurs de Baseball regroupant plusieurs plusieurs indicateurs de performance, salaire, carrière et équipes.

Problématique et objectif

La variable salaire est de l'année 1987 et la majorité des autres variables est de l'année 1986. Il est donc intéressant de voir l'impact qu'ont eu les différentes performances des joueurs en 1986 sur leur salaire en 1987. Nous essaierons d'établir un lien entre performances et salaire des joueurs de Baseball de notre dataset. (Est-ce que les salaires sont mis à jour chaque année ?)

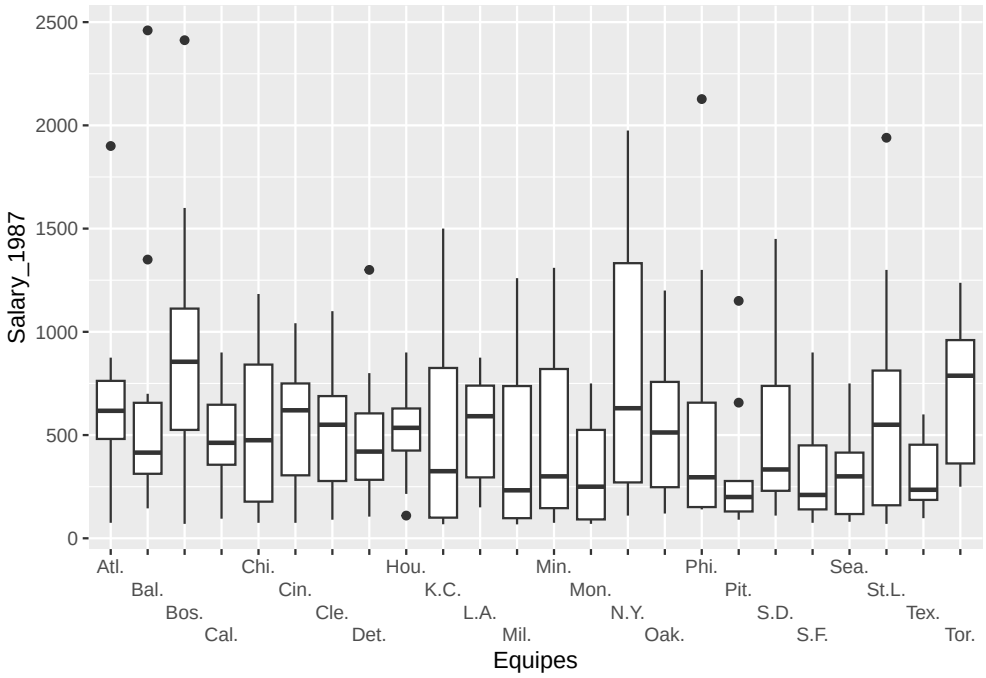
Dans quelle mesure la performance d'un joueur explique son salaire ?

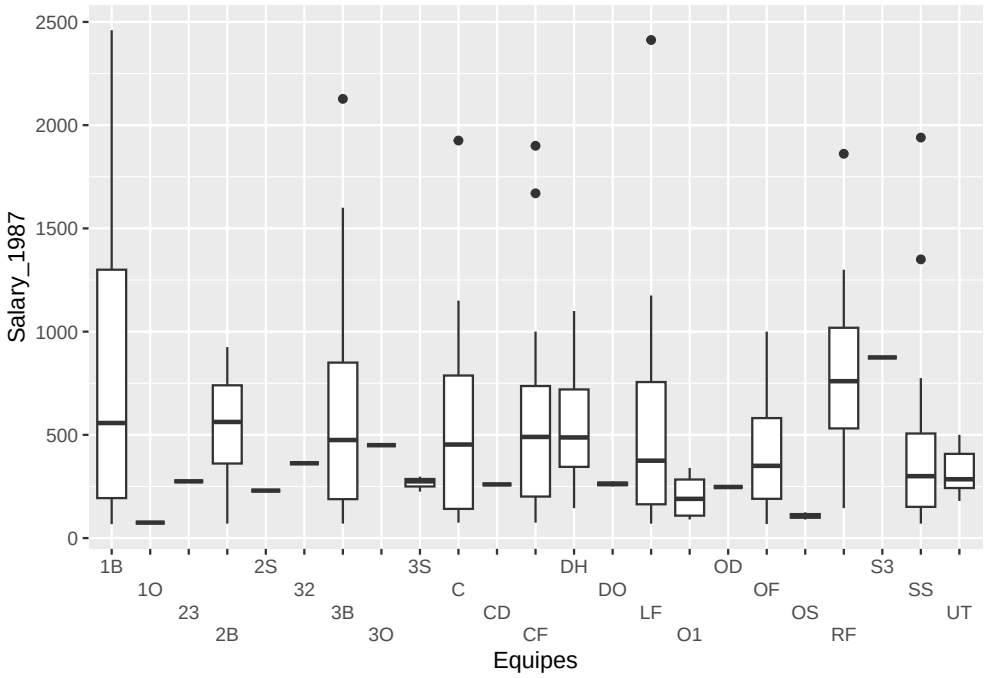
I - Analyse descriptive des données

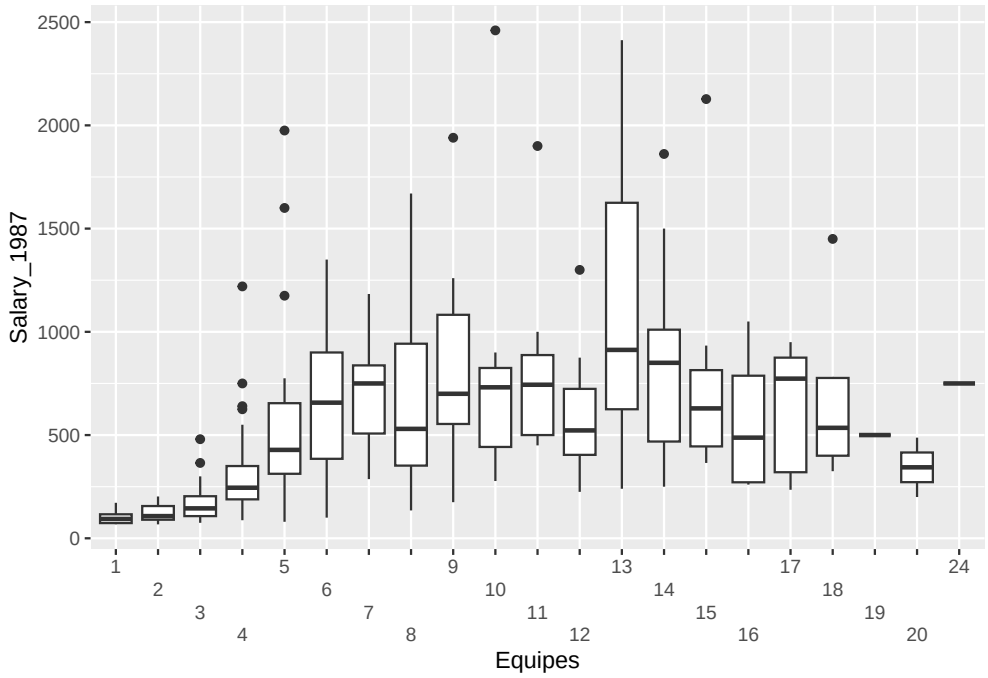
A - Analyse du jeu de données

Après importation du jeu de données, voici un tableau résumant les variables quantitatives :

Salaires en fonction de l'équipe

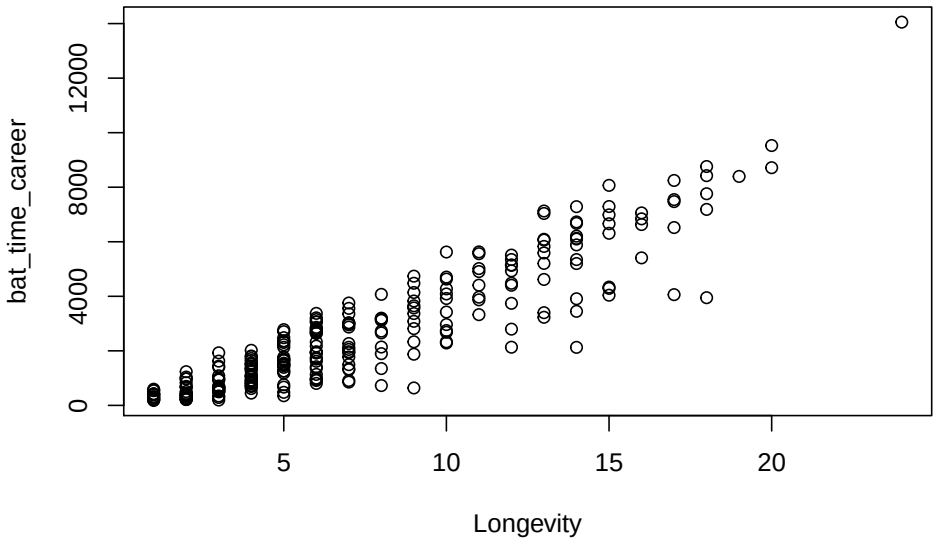

Salary_1987
2500
2000
1500
1000
500
0
Atl.
Bal.
Bos.
Cal.
Chi.
Cin.
Cle.
Det.
Hou.
K.C.
L.A.
Mil.
Min.
Mon.
N.Y.
Oak.
Phi.
Pit.
S.D.
S.F.
Sea.
St.L.
Tex.
Tor.
Equipes

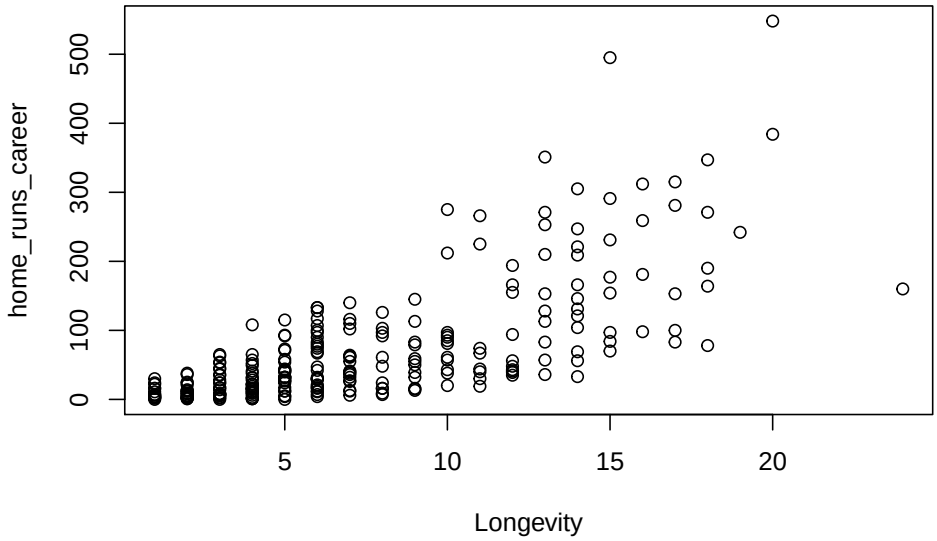

Salary_1987
2500
2000
1500
1000
500
0
1B
1O
23
2B
2S
32
3B
3O
3S
C
CD
CF
DH
DO
LF
O1
OD
OF
OS
RF
S3
SS
UT
Equipes

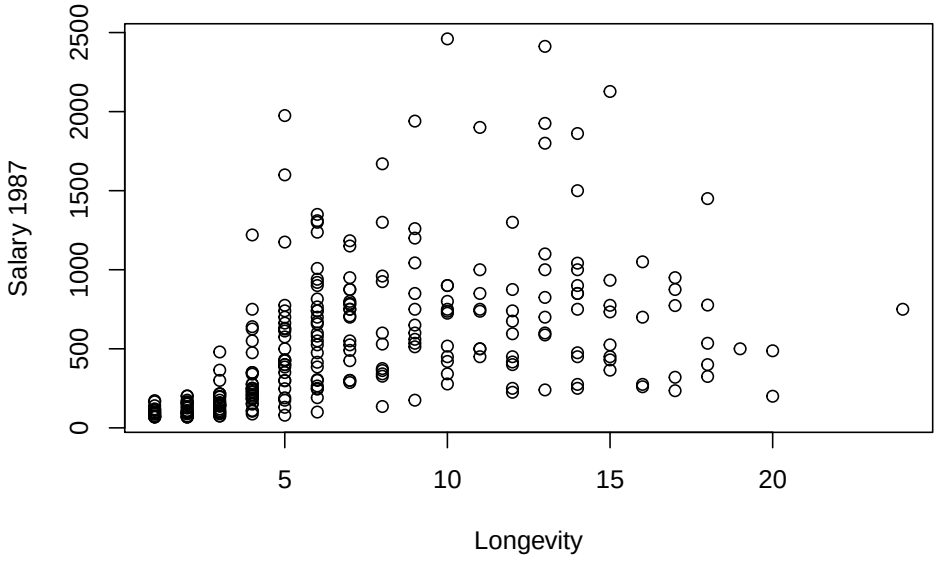


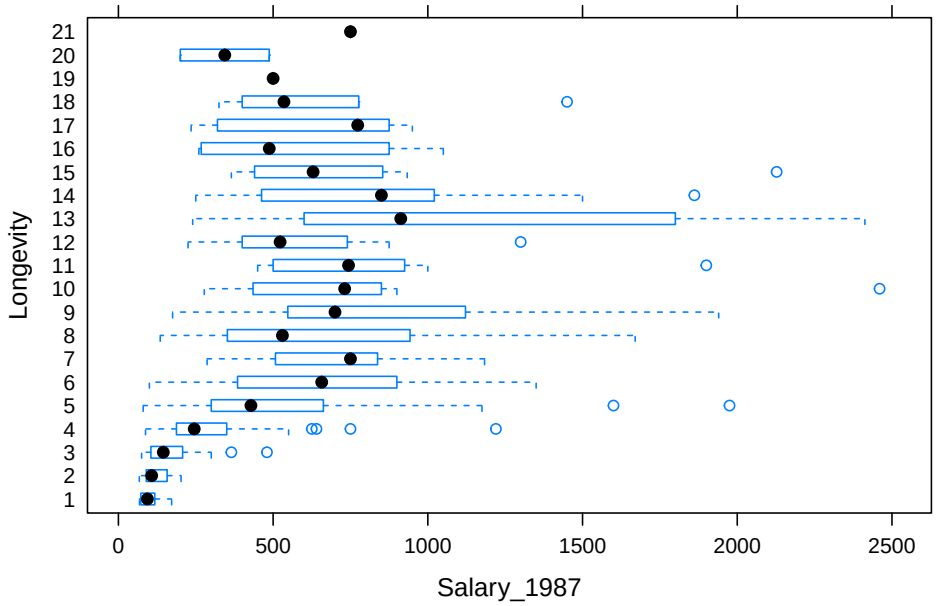
B - Analyse univariée

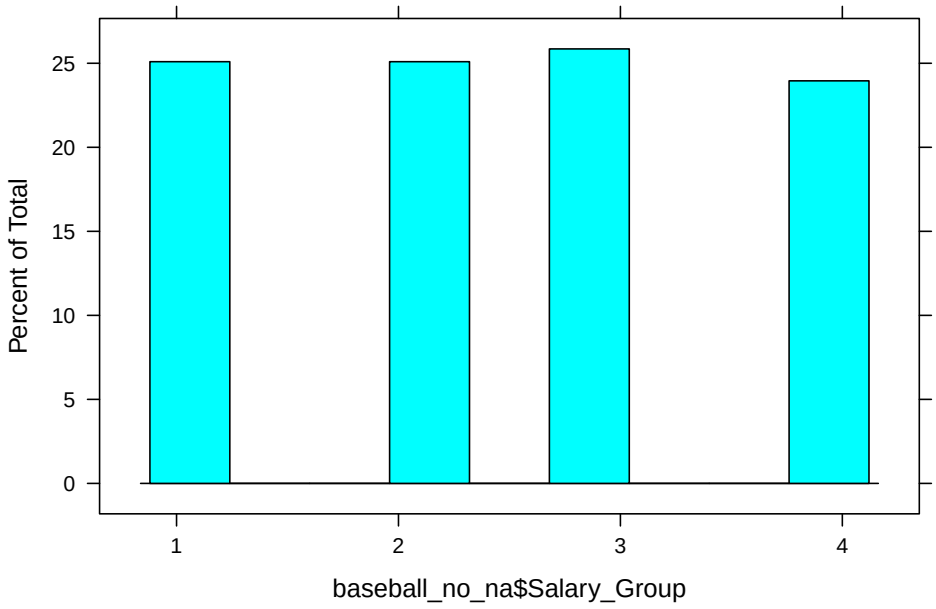
Il peut être intéressant de regarder



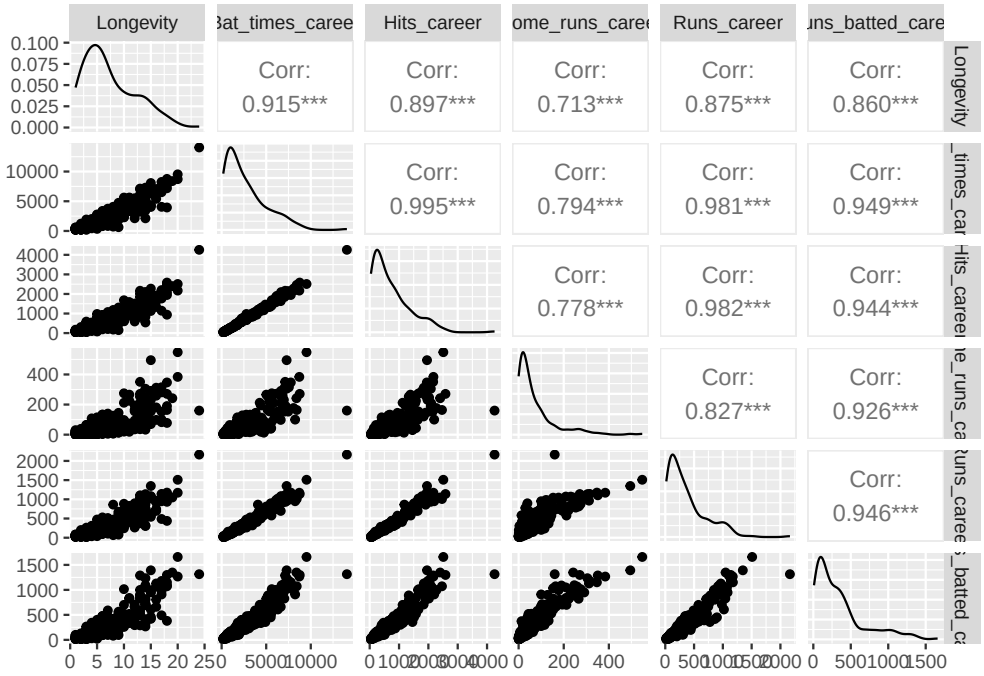

home_runs_career
0
100
200
300
400
500
5
10
15
20
Longevity


Salary 1987
0
500
1000
1500
2000
2500
5
10
15
20
Longevity


Longevity
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
500
1000
1500
2000
2500
Salary_1987


Percent of Total
25
20
15
10
5
0
1
2
3
4
baseball_no_na$Salary_Group

C - Analyse bivariée



II - Etude groupée

A - Anova

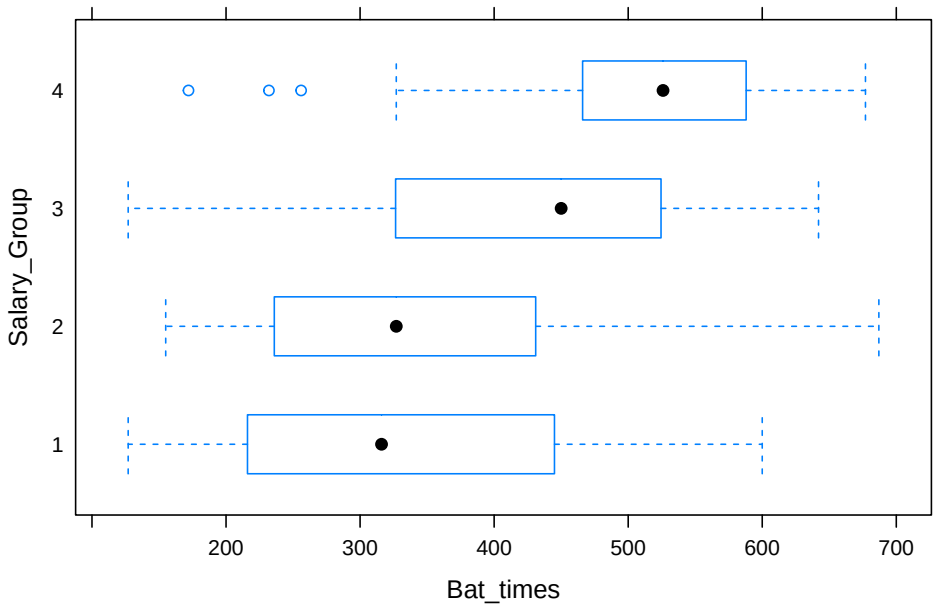
Anova

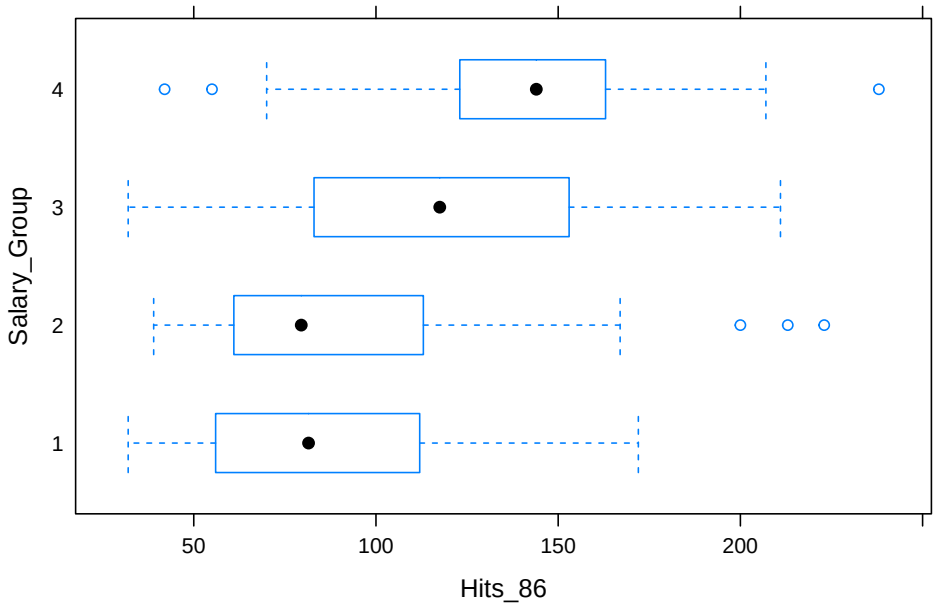
```
aov.res <- aov(Salary_1987 ~ Longevity, data = baseball)
summary(aov.res)
```

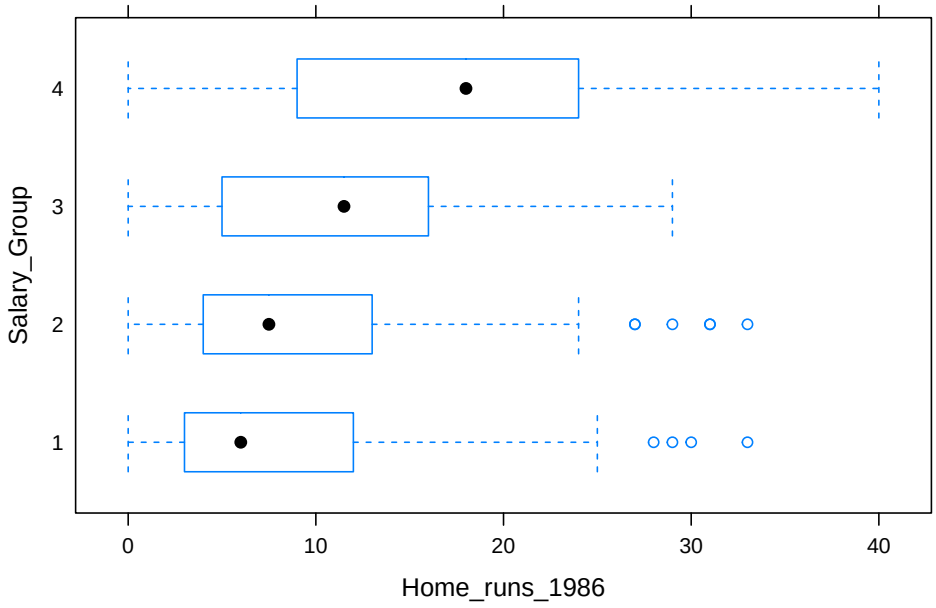
```
             Df   Sum Sq  Mean Sq F value   Pr(>F)
Longevity     1 10440243 10440243   63.95 4.16e-14 ***
Residuals   261 42609462   163255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Puis

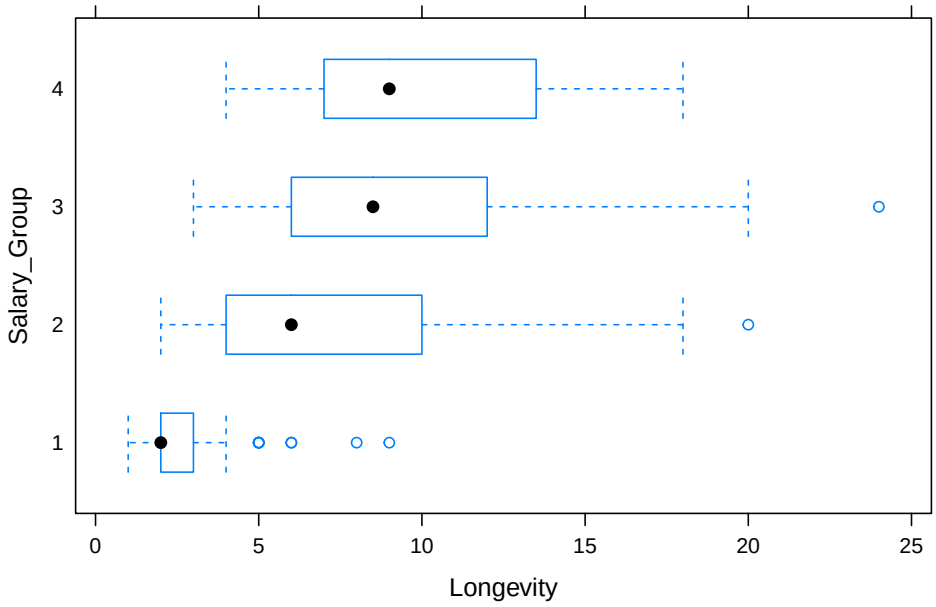
```
              Df  Sum Sq Mean Sq F value   Pr(>F)
Salary_Group   1 1059410 1059410   64.74 3.01e-14 ***
Residuals    261 4270767   16363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Salary_Group
4
3
2
1
200
300
400
500
600
700
Bat_times


Salary_Group
4
3
2
1
50
100
150
200
Hits_86

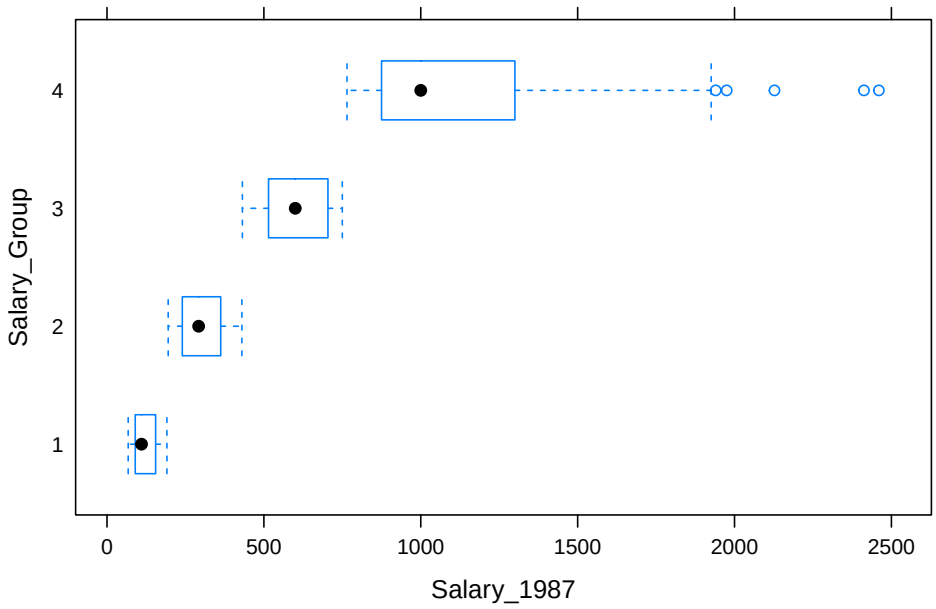



```
              Df Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1     35   35.03   0.505  0.478
Residuals    130   9014   69.34
```

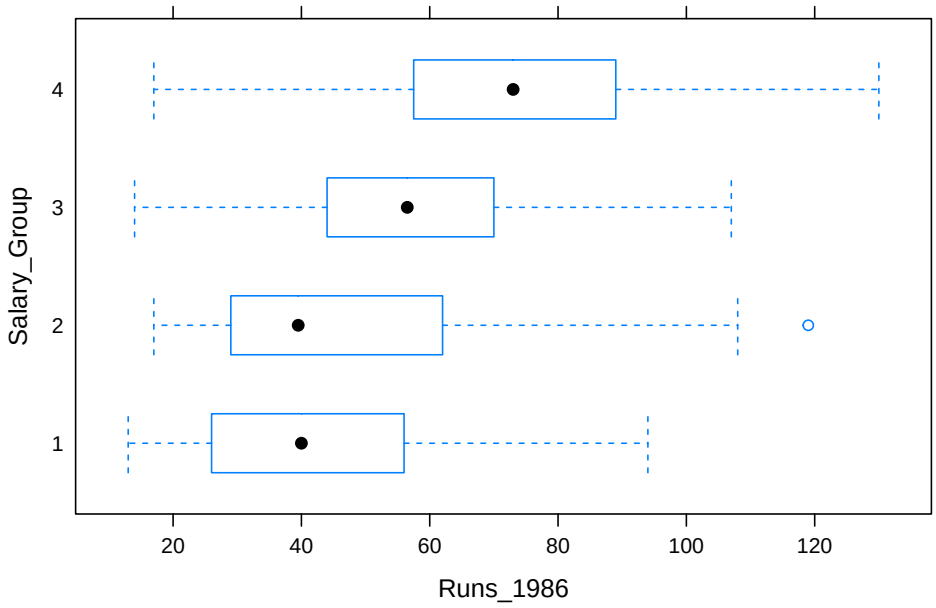



```
              Df Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1 117473  117473    78.7 <2e-16 ***
Residuals    261 389604    1493
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

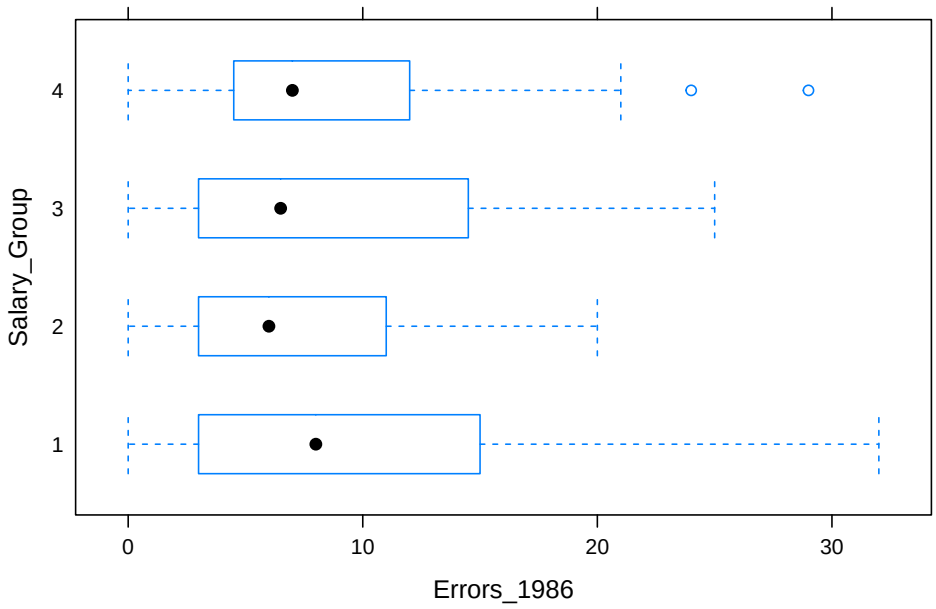
```
              Df Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1   1140    1140   0.758  0.386
Residuals    130 195587    1504
```




```
              Df  Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1 1061086 1061086   314.1 <2e-16 ***
Residuals    130  439110    3378
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

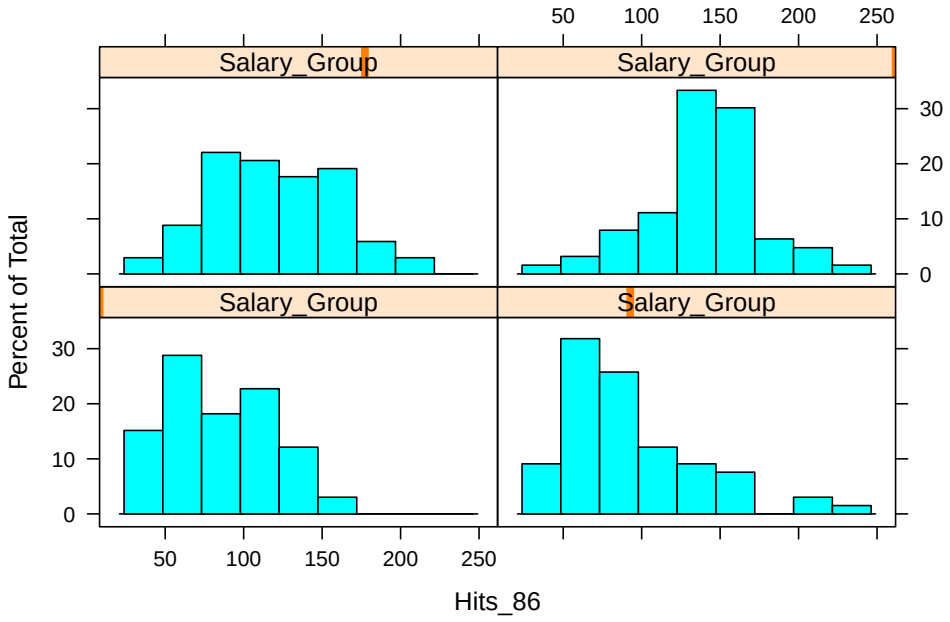


```
              Df Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1    810   810.1   1.513  0.221
Residuals    130  69604   535.4
```




```
              Df Sum Sq Mean Sq F value Pr(>F)
Salary_Group   1      6    6.23   0.142  0.706
Residuals    261  11429   43.79
```

B - Histogrammes



C - Tests de student

```
    Welch Two Sample t-test

data:  Hits_86 by Salary_Group
t = -0.87066, df = 124.88, p-value = 0.3856
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.24224   7.48466
sample estimates:
mean in group 1 mean in group 2
       86.19697        92.07576
```

III - Regressions linéaires

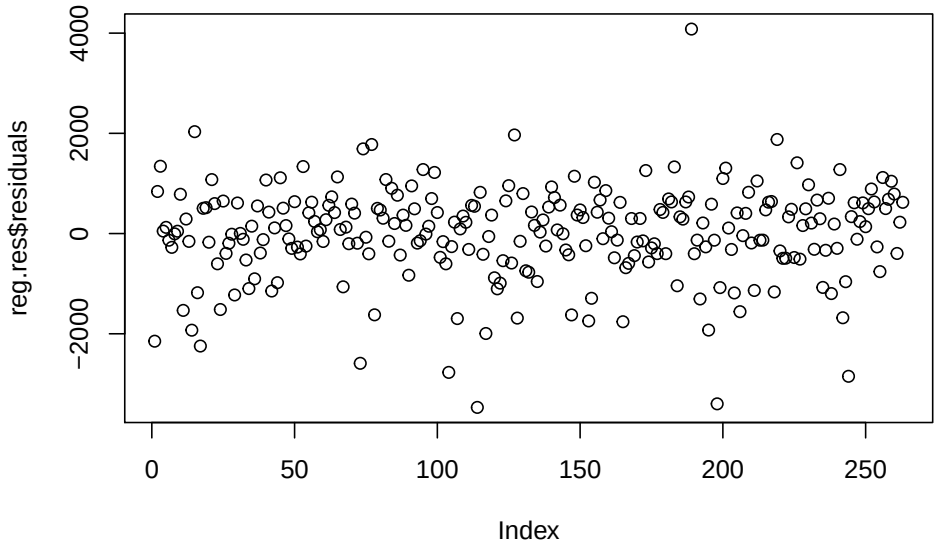
A - Simple

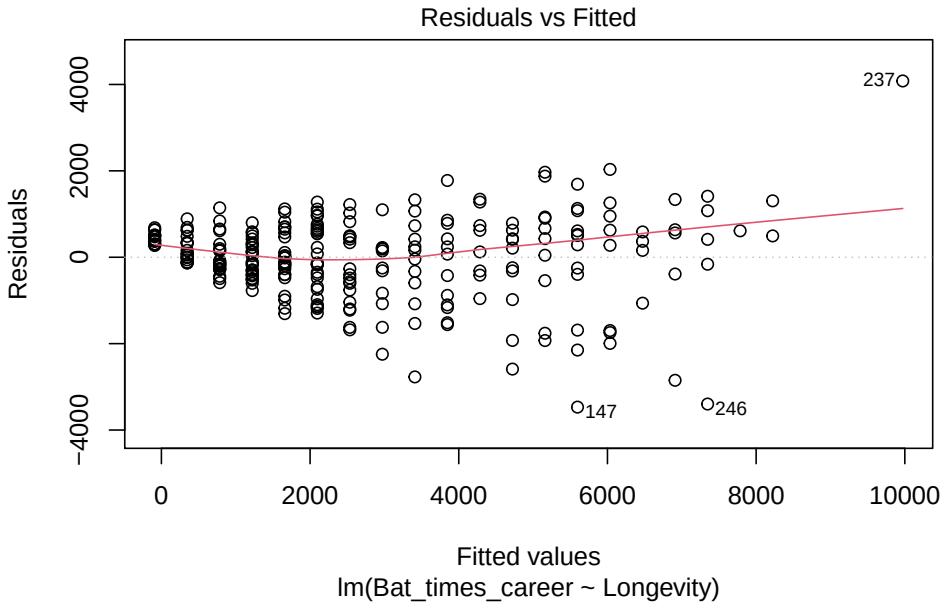
cas simple

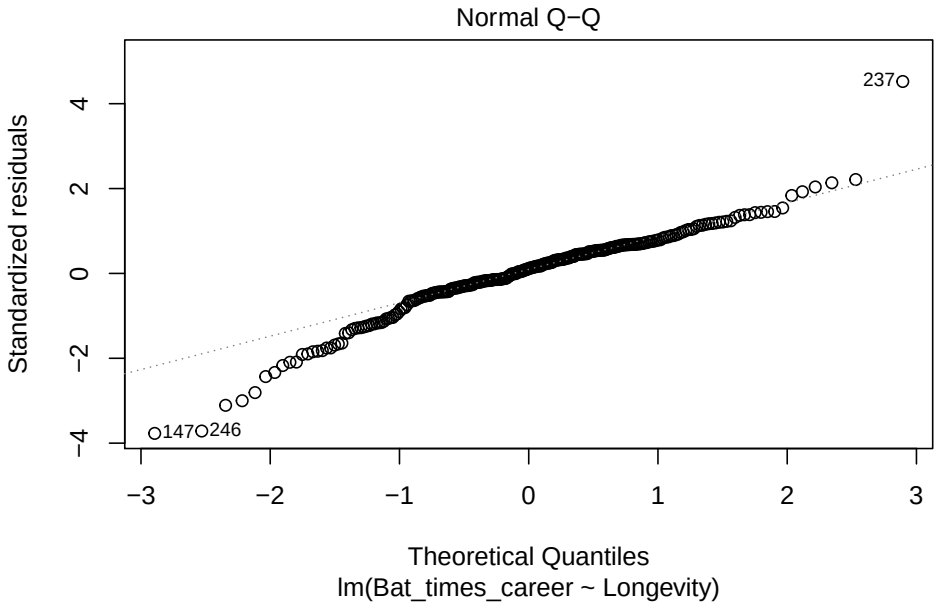
```
reg.res <- lm(Bat_times_career ~ Longevity, data = baseball)
summary(reg.res)
```

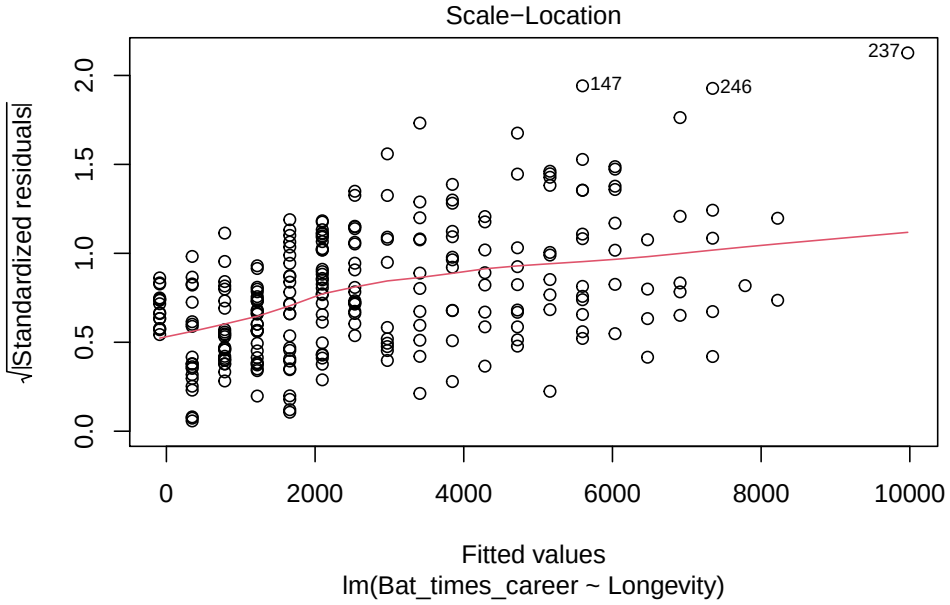
```
Call:
lm(formula = Bat_times_career ~ Longevity, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-3469.2  -402.0   110.5   576.4  4080.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -528.10     104.95  -5.032 9.04e-07 ***
Longevity     437.52      11.93  36.660  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 925.4 on 261 degrees of freedom
Multiple R-squared:  0.8374,    Adjusted R-squared:  0.8368
F-statistic:  1344 on 1 and 261 DF,  p-value: < 2.2e-16
```

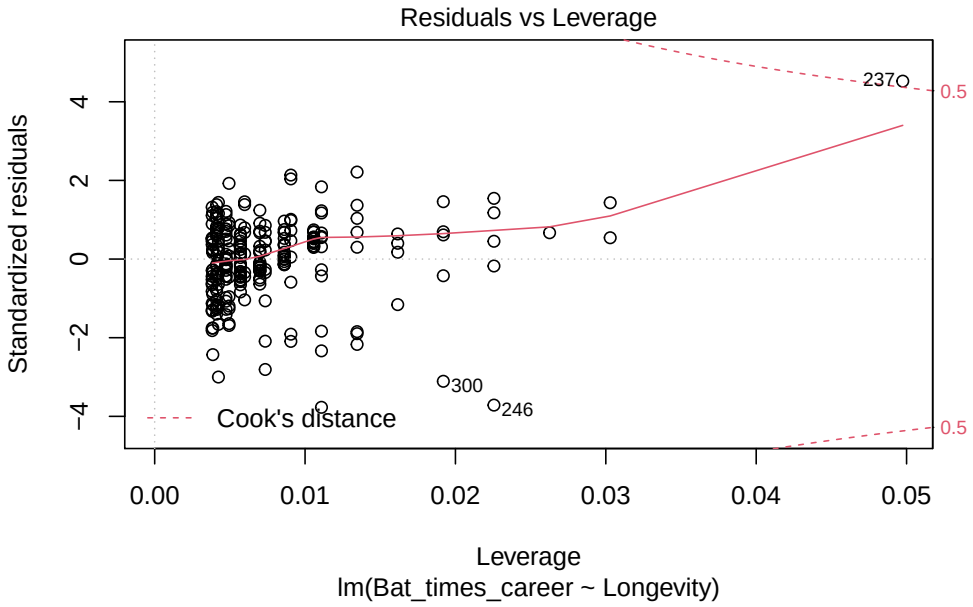
résidus


reg.res$residuals
4000
2000
0
−2000
0
50
100
150
200
250
Index


Residuals vs Fitted
Residuals
4000
2000
0
−4000
237
147
246
0
2000
4000
6000
8000
10000
Fitted values
lm(Bat_times_career ~ Longevity)

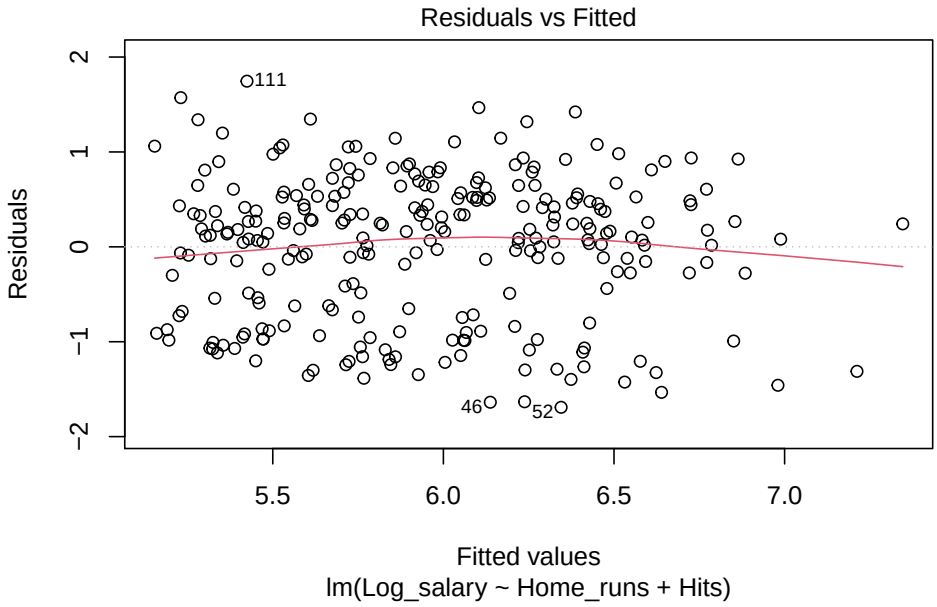

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Bat_times_career ~ Longevity)
237
147
246

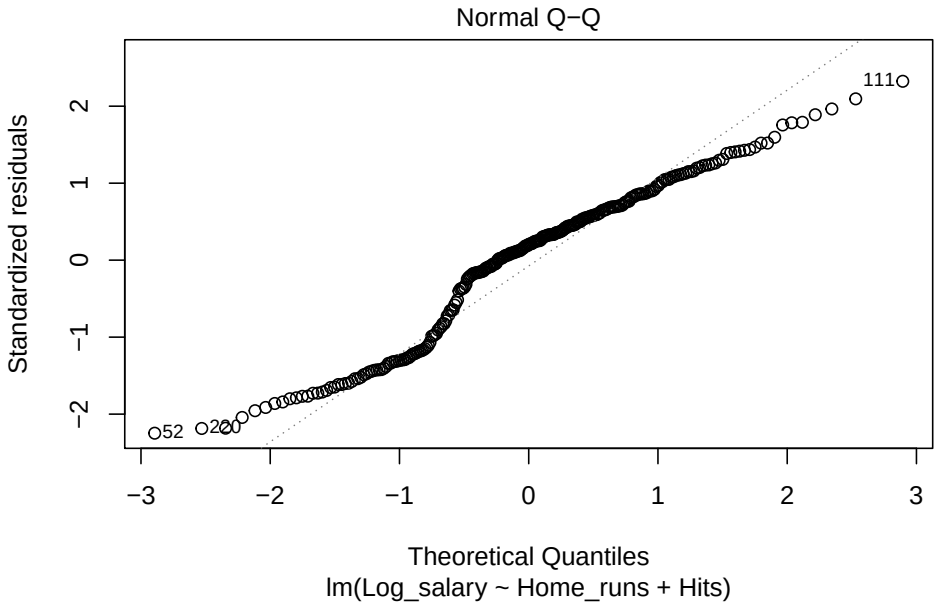

Scale−Location
√|Standardized residuals|
Fitted values
lm(Bat_times_career ~ Longevity)
237
147
246

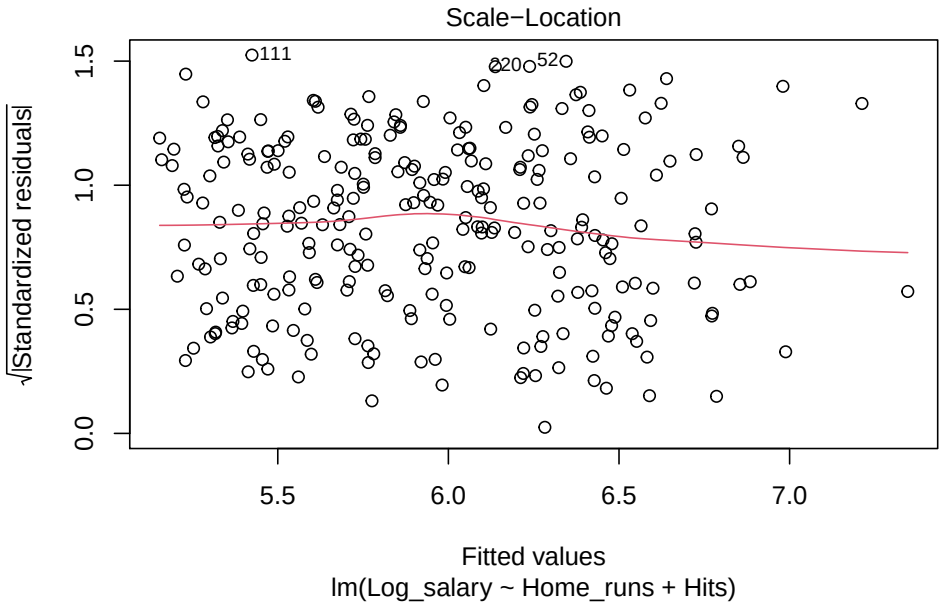


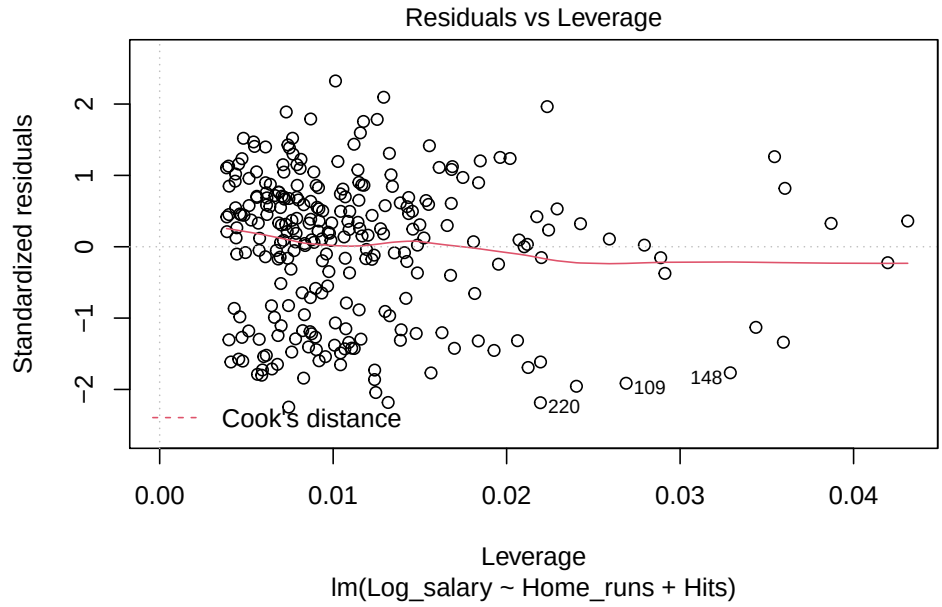
B - Multiple

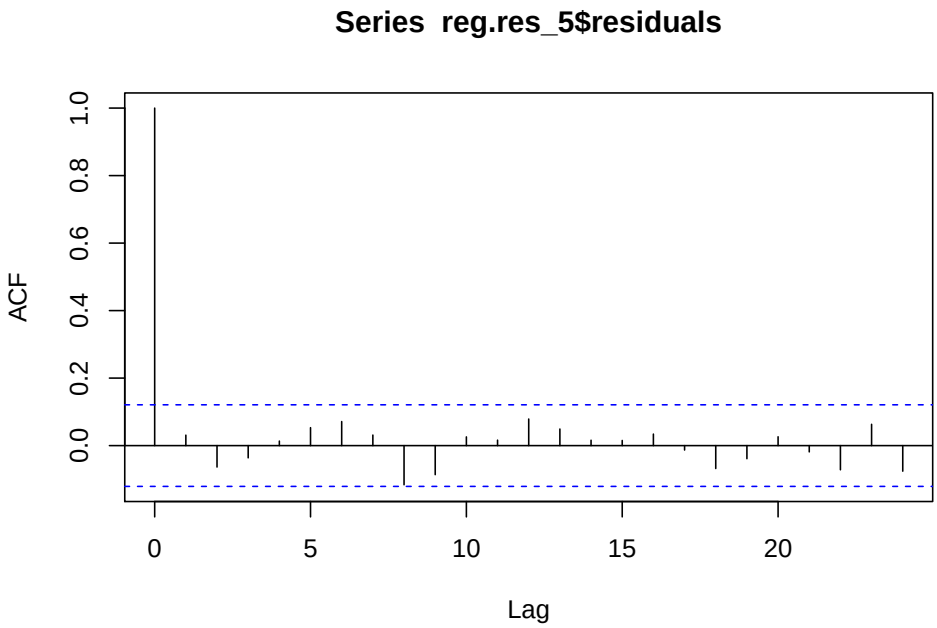
```
Call:
lm(formula = Log_salary ~ Home_runs + Hits)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6911 -0.6371  0.1529  0.5227  1.7458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.815805   0.124806  38.586  < 2e-16 ***
Home_runs   0.012952   0.006200   2.089   0.0377 *
Hits        0.008945   0.001244   7.192 6.81e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7552 on 260 degrees of freedom
Multiple R-squared:  0.2752,    Adjusted R-squared:  0.2697
F-statistic: 49.37 on 2 and 260 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
lm(Log_salary ~ Home_runs + Hits)
111
46
52
5.5
6.0
6.5
7.0
−2
−1
0
1
2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Log_salary ~ Home_runs + Hits)
111
52
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
111
220
52
0.0
0.5
1.0
1.5
5.5
6.0
6.5
7.0
Fitted values
lm(Log_salary ~ Home_runs + Hits)


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
Cook's distance
220
109
148
0.00
0.01
0.02
0.03
0.04
Leverage
lm(Log_salary ~ Home_runs + Hits)



Conclusion

Annexe