Projet d'analyse de données - Baseball

Hicham, Ryad, Elyass et Hongxin

17 décembre 2022

Contents

Introduction

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

Problématique et objectif

La variable salaire est de l'année 1987 les autres variables sont de l'année 1986 ou s'étalent sur la carrière des joueurs. 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 ?.. Nous n'avons que ces données.)

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

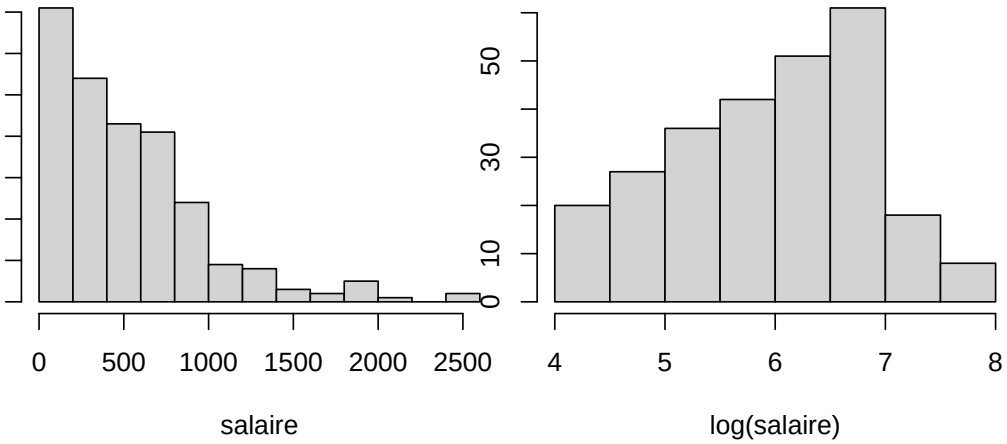
I - Analyse descriptive des données

A - Analyse du jeu de données

Le jeu de données possède 28 variables dont 7 variables qualitatives. Après importation et retrait des lignes avec des valeurs manquantes, on obtient un tableau de 263 joueurs. Ce sont tous des hitters.

B - Analyse univariée

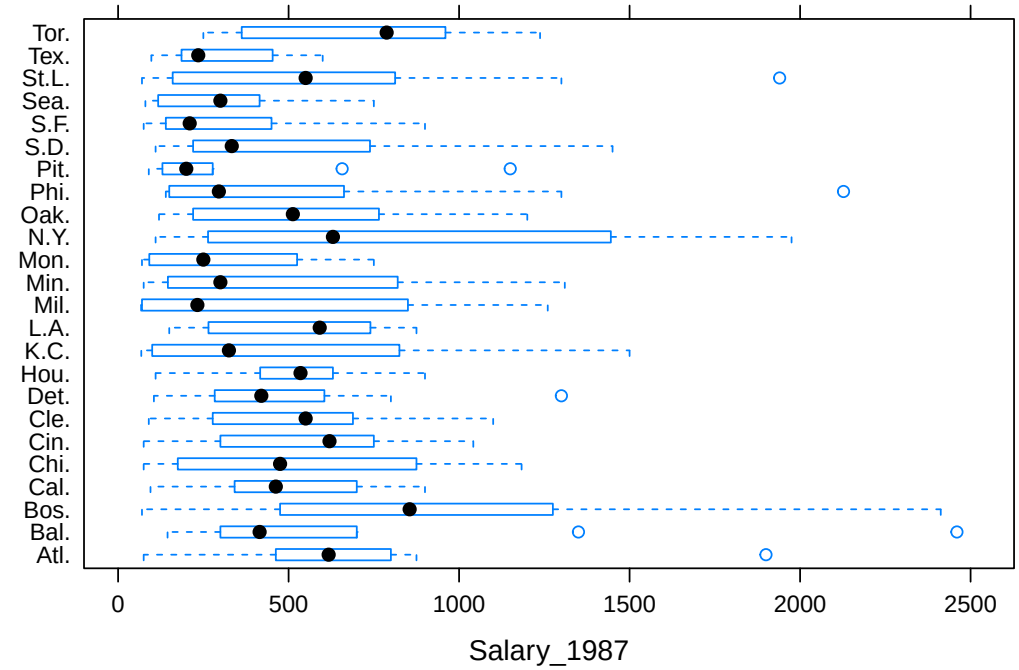
L'histogramme des salaires nous permet de visualiser l'étendue et la distribution de cette variable. La distribution ressemblant à une distribution de type exponentielle, on regarde le Log du salaire pour avoir une distribution symétrique, réduire la variance de l'échantillon et diminuer la grande différence entre les petits et grands salaires. Surtout, c'est pour avoir plus de chances d'expliquer cette variable à l'aide d'un modèle linéaire en fonction des autres.




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
67.5   193.0   430.0   542.2   750.0  2460.0
```

Ajoutons les quartiles des salaires : les 1ers et 2e quartiles seront les petits salaires, et les 3e et 4e quartiles les grands salaires.

Affichons les salaires des joueurs en fonction de leur équipe :



On observe que l'équipe d'appartenance est discriminante pour le salaire. En particulier, il est évident qu'une anova nous donnera une influence de l'équipe sur le salaire. Même si les équipes contiennent moins de 30 joueurs et qu'on a pas de normalité asymptotique.

Puis leur salaire en fonction de leur ancienneté :

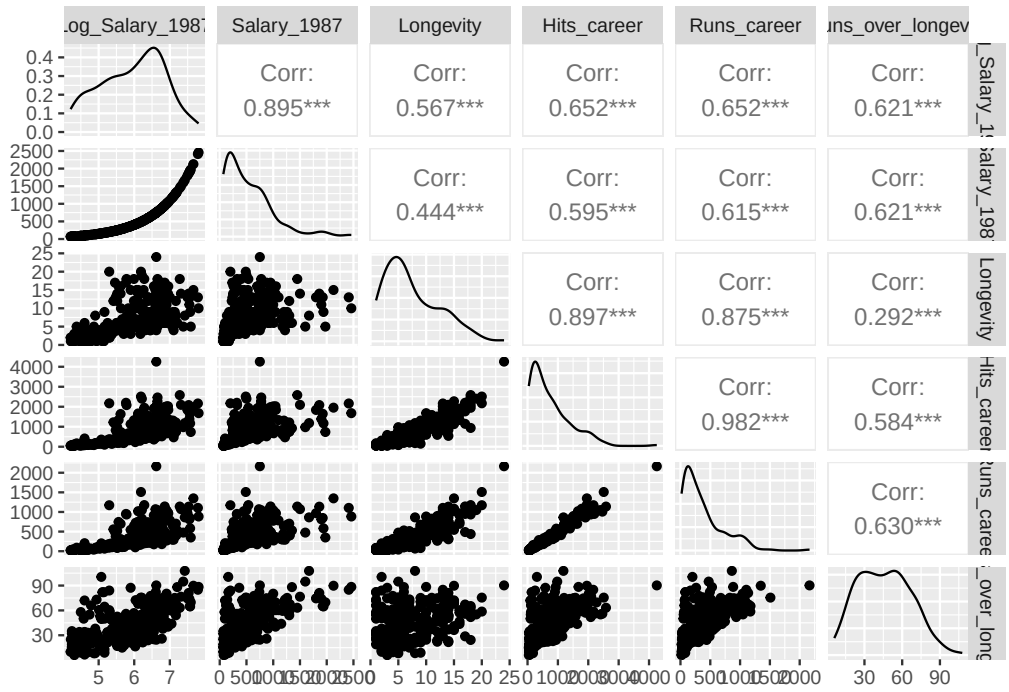
Nous constatons l'existence de trois phases d'évolution de salaires par rapport aux années expériences professionnelles:

- Phase 1: [1,5] c'est le début de carrières des joueurs de Baseball avec des salaires faibles mais qui augmentent avec le temps. L'émergence de nouvelles stars avec des salaires remarquables.
- Phase 2: [6,13] c'est la phase de la maturité professionnelles où des joueurs se différencient par rapport à la médiane, c'est l'age d'or des joueurs. Ils ont touché des salaires importants.
- Phase 3: [14,24] c'est la fin des carrières, nous observons que des joueurs démissionnent à partir de l'année 18, les salaires en parallèles diminuent. Des cas de figures exceptionnels restent toujours sur le marché et réussissent à garder leur salaires intéressants.

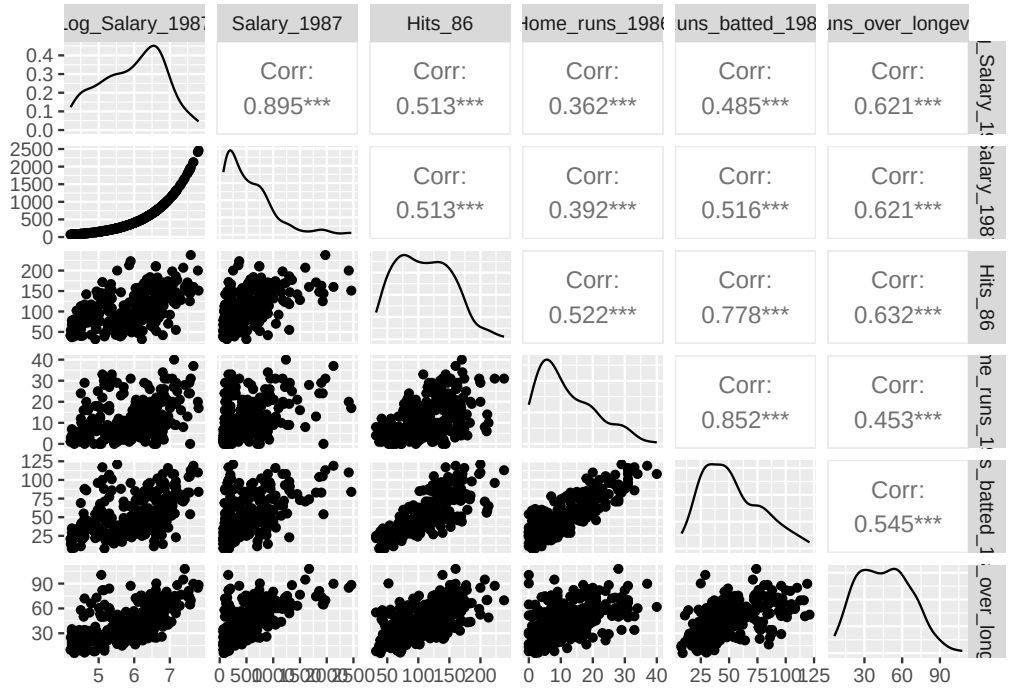
La longévité est bien discriminante sur le salaire.

C - Analyse bivariée

On va regarder deux tableaux de corrélations : les corrélations entre les variables qui s'étendent sur la carrière, et celles qui s'étendent sur les années. On va chercher des liens dans les données afin de mieux expliquer le salaire.



Deux variables ont été ajoutées au dataframe : Log_Salary et Runs_over_longevity. On remarque plusieurs tendances, en particulier des corrélation évidentes entre le le nombre de Hits_career et le nombre de Runs_career.

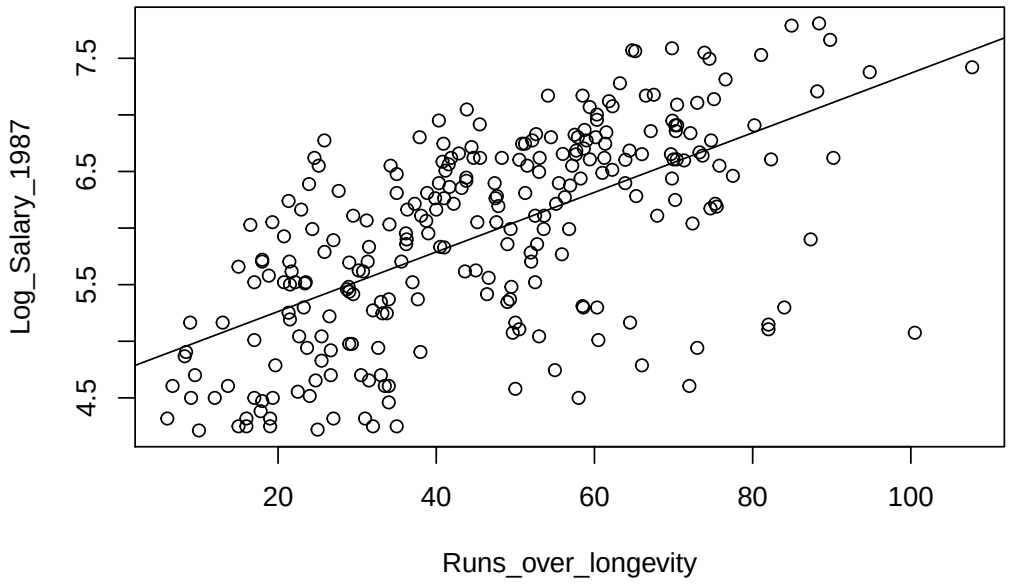


Pour ce qui est du salaire, le tableau suggère un lien linéaire entre le Log_salaire et Runs_over_longevity. Le nuage de points semble + centré autour d'une éventuelle droite que celui du salaire, qui serait + utile pour de la classification.

II - Régressions linéaires

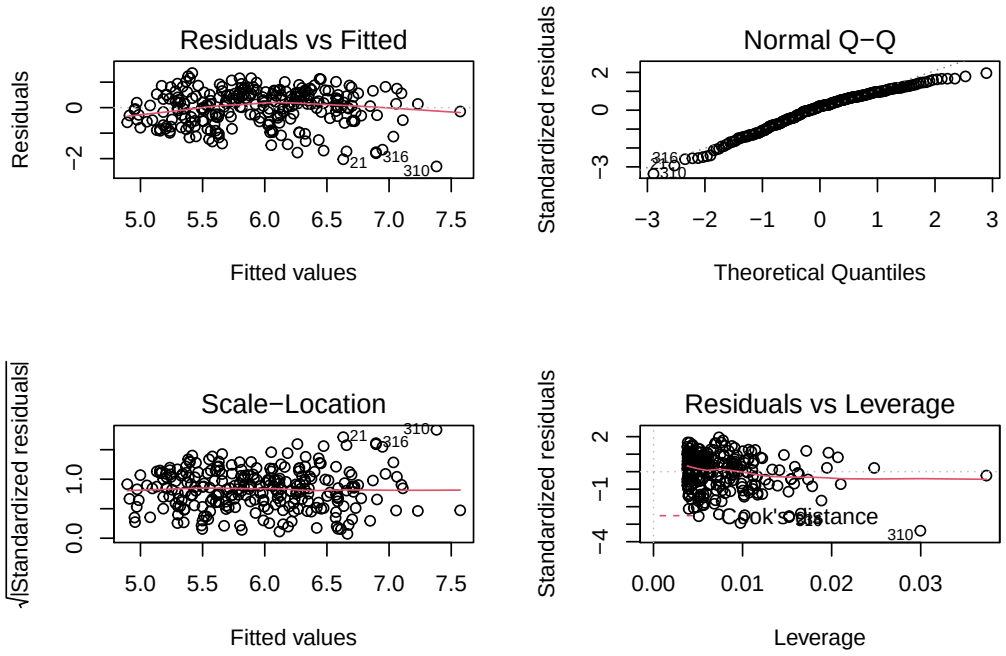
A - Régression simple

Comme l'a suggéré le tableau des corrélations, on regarde le lien linéaire entre Log_Salary_1987 et Runs_career/Longevity = Runs_over_longevity




```
Call:
lm(formula = Log_Salary_1987 ~ Runs_over_longevity, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3063 -0.4525  0.1429  0.5048  1.3575

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.736136   0.103777   45.64   <2e-16 ***
Runs_over_longevity 0.026322   0.002058   12.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6942 on 261 degrees of freedom
Multiple R-squared:  0.3853,    Adjusted R-squared:  0.383
F-statistic: 163.6 on 1 and 261 DF,  p-value: < 2.2e-16
```

Les statistiques de tests sont claires, on rejette l'hypothèse que les coefficients de régression sont nuls.

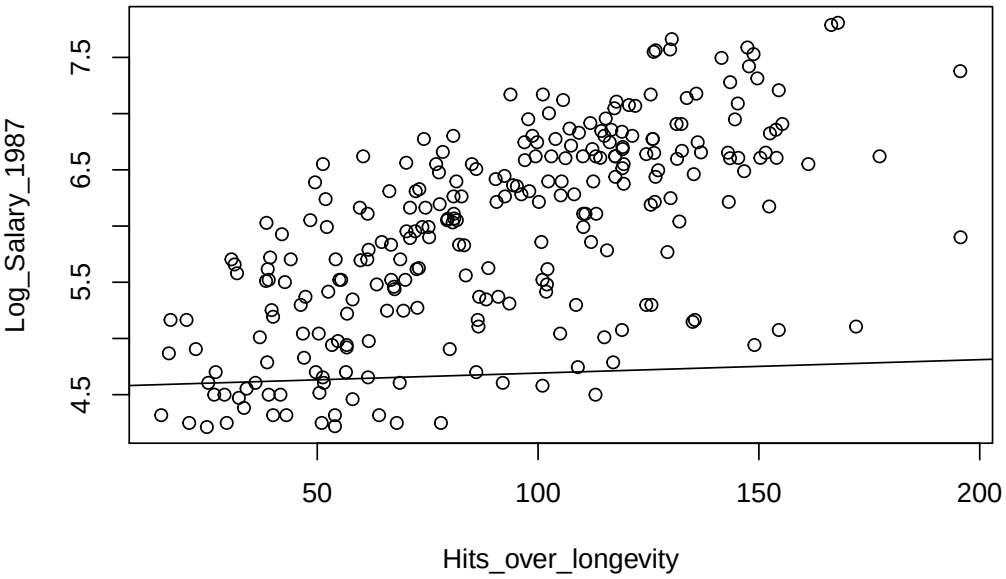


On a bien la normalité des résidus d'après le QQ-plot. Pour les distances de Cook, certaines ont l'air élevées, mais elles ne dépassent pas le contour en poitillés qui n'est pas visible. Il y a donc des valeurs aberrantes dans cette régression mais elles ne sont pas des points leviers.

B - Régressions multiples

On va essayer d'expliquer le salaire à partir des variables Runs_career et Hits_careers. Mais les joueurs ne sont pas au même stade de leur carrière. Alors on les divise par Longevity pour avoir des performances moyennes.

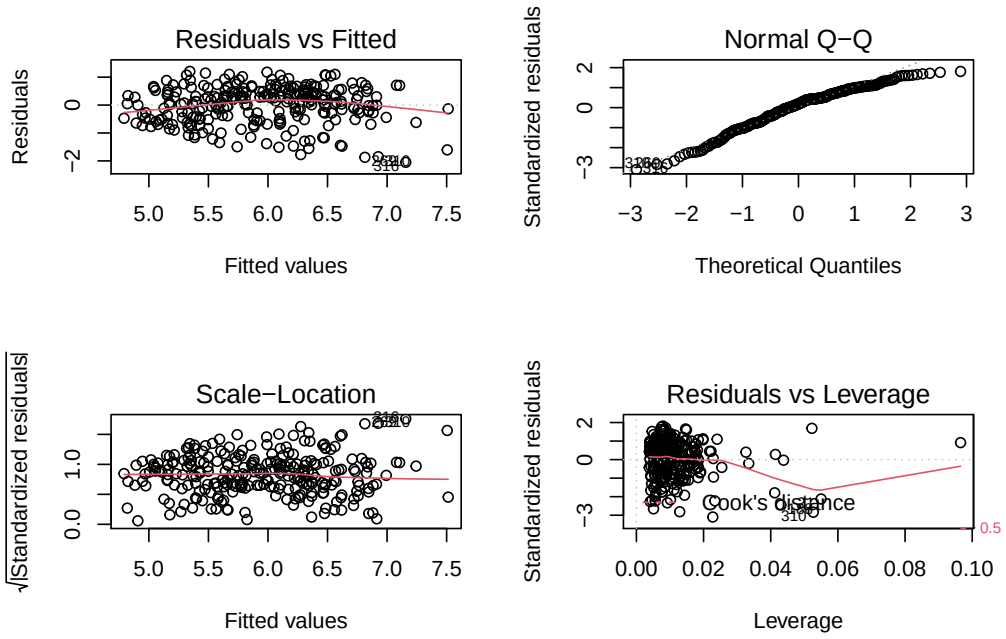
```
Warning in abline(reg.multiple.longevity): only using the first two of 3
regression coefficients
```



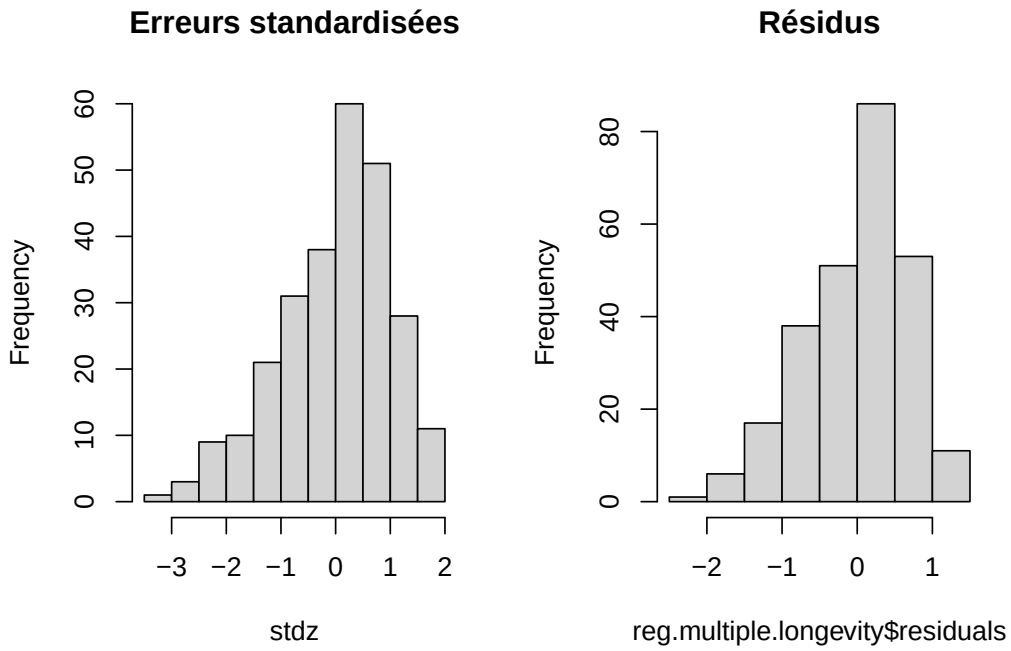

```
Call:
lm(formula = Log_Salary_1987 ~ Runs_over_longevity + Hits_over_longevity,
```

```
    data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0508 -0.4613  0.1293  0.4861  1.2073

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.572311   0.106500  42.933  < 2e-16 ***
Runs_over_longevity 0.001195   0.005909   0.202     0.84
Hits_over_longevity 0.014456   0.003202   4.515 9.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6697 on 260 degrees of freedom
Multiple R-squared:   0.43, Adjusted R-squared:  0.4256
F-statistic: 98.07 on 2 and 260 DF,  p-value: < 2.2e-16
```

De même, ici on peut rejetter les hypothèses sur les coefficients de régression nuls. Pour l'effet cumulé de Runs et Hits, on peut rejetter la nécessité de garder ce



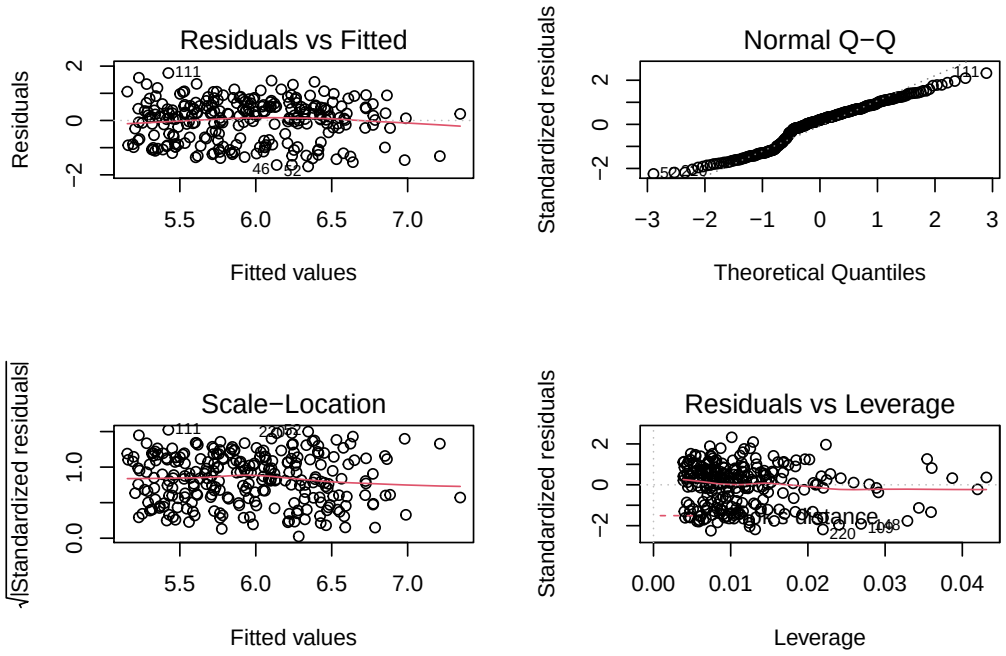
Les résidus sont toujours bien normaux. Pour les valeurs aberrantes, on observe bien un point levier qui est Tim Teufel avec un salaire de 277.5 qui as un



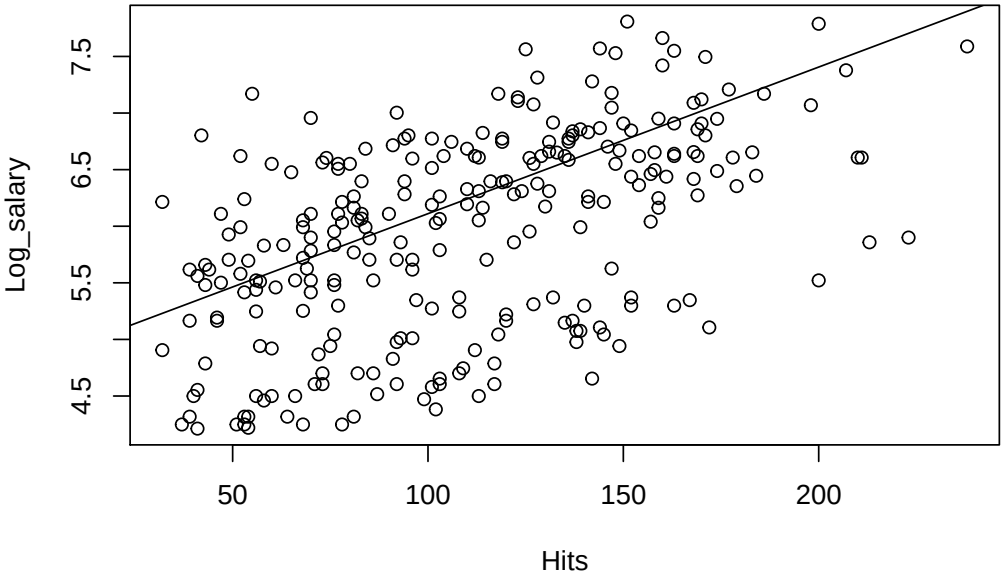
expliquer distribution résidus standardisés et résidus

On refait la même régression, sauf qu'on prend uniquement les variables Hits_86 et Runs_1986

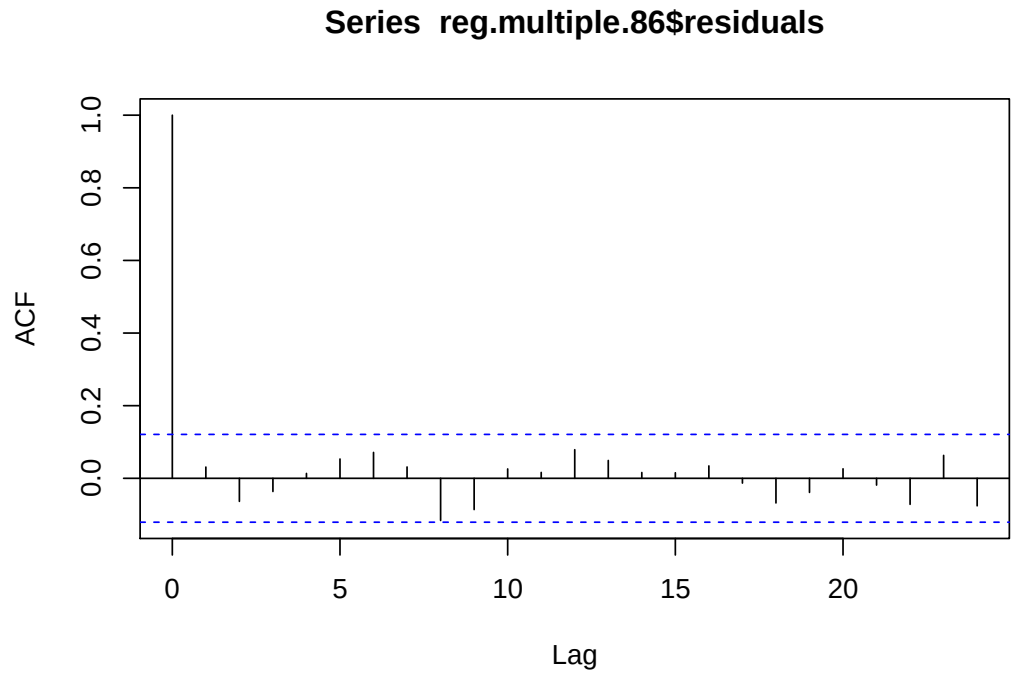
```
Call:
lm(formula = Log_salary ~ Home_runs + Hits)

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
```




```
Warning in abline(reg.multiple.86): only using the first two of 3 regression
coefficients
```



Par rapport à la première régression multiple, on voit que les Hits de l'année dernière expliquent mieux le salaire que les Hits moyens sur la carrière.



expliquer absence d'autocorrelation dans les erreurs.

III - Etude groupée

A - Anova

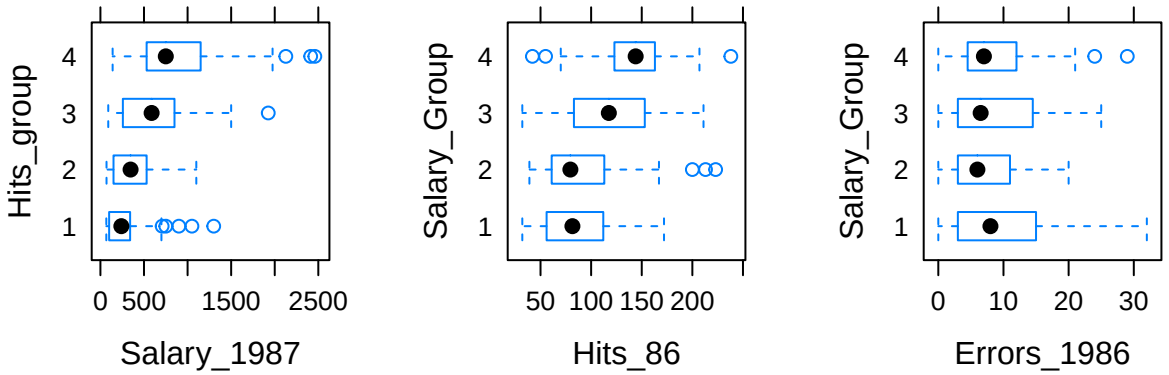
On a déjà vu dans la partie I que la longévité était discriminante pour le salaire. On le vérifie avec une anova, même si l'hypothèse de normalité n'est pas vérifiée. Certaines valeurs de longévités de contiennent que quelques joueurs.

```
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
```

On regroupe alors les salaires en 4 groupes des 4 quantiles de la distribution afin de réaliser une nouvelle anova.

```
              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
```



On distingue clairement une tendance entre les 2 premiers groupes de salaires et les deux derniers. C'est à dire par rapport à la médiane des salaires. Être payé au dessus de la médiane est nécessaire pour être un meilleur hitteur, modulo quelques exceptions. La réciproque n'est pas vraie. En faisant 4 groupes de Hitteurs, être au dessus de la médiane des hits en 86 n'est pas nécessaire un salaire élevé : il y a des mauvais hitteurs dans le premier quartile qui sont bien payés qui se démarquent.

```
              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
```

Explication de anova hits

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

Explication de anova hits lower

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

Explication de anova home runs

B - 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
```

Explications test de student.

Conclusion

On conclut