Analyse de l'influence des performances des joueurs de NBA sur leur rémunération

Samuel BALLU Paul VALLEE

2023/2024

Introduction

L'intersection entre la performance des joueurs de la NBA et leur rémunération a toujours été un sujet captivant. Notre projet se plonge dans cette dynamique, examinant les données de 1996 à 2017 pour comprendre comment les performances individuelles influent sur les contrats et les salaires de nos joueurs. À travers une approche Économétrique, nous explorons les liens entre les perfomances sur le terrain et la compensation financière des joueurs. Dans ce contexte nous nous posons la question suivante :

Dans quelle mesure les performances individuelles des joueurs de basket influent-elles sur leur salaire?

Cette analyse s'inscrit dans le contexte plus large des débats actuels sur les salaires des athlètes professionnels, offrant des perspectives significatives pour les passionnés de basketball, les analystes économiques, et les acteurs du monde sportif.

1 Présentation des jeux de données

1.1 Explication générale

Les données utilisées dans ce projet proviennent de deux sources différentes.La source initiale provient d'une base de données de la NBA contenant les statistiques de chaque joueur de la ligue de 1996 à 2017. La seconde est une base de données inscrivant les salaires de chaque joueur de la ligue de 1990 à 2017. Nous avons associé nos individus, les joueurs de basket, en fonction de leurs équipes et des salaires perçus pour chaque saison, en se basant sur leurs performances de l'année précédente. Cette approche découle de notre hypothèse selon laquelle les salaires sont déterminés uniquement en fonction des performances antérieures à la saison en cours, car nous considérons que les salaires sont réévalués entre les saisons en se fondant sur les anciennes performances. Notre base de données comprendra au final 11 145 observations et 25 variable sur 2463 joueurs.

1.2 Dictionnaire des variables

Variable	Description	Unité
Player_Name	Nom du joueur	-
Season_Start	Année en cours	Date
Team	Acronyme du nom de l'équipe	-
age	Age du joueur	-
player_height	Taille du joueur	cm
player_weight	Poids du joueur	kg
college	Université à laquelle le joueur était inscrit	-
country	Pays d'origine du joueur	-
draft_year	Année de draft d'apparition du joueur en NBA	-
draft_round	Tour auquel le joueur a été drafté	-
draft_number	Position à laquelle le joueur a été drafté	-
gp	Nombre de match joués par le joueur	-
pts	Nombre de points moyen marqué	-
reb	Nombre de rebonds moyen	-
ast	Nombre de passes décisives	-
net_raiting	Différence moyenne entre le nombre de points	-
oreb_pct	Rebonds offensif pris par le joueur	%
dreb_pct	Rebonds défensif pris par le joueur	%
usg_pct	Possession de la balle de l'équipe	%
ts_pct	Pourcentage réel de réussite du joueur	%
ast_pct	Passes décisives du joueur	%
Salary	Salaire du joueur	$ Prix constant de 2010

1.3 Importation des données

Pour pouvoir faire nos analyses, nous avons trouvées nos deonnéees sur internet. Le premier contient les statistqiues des joueurs de NBA, le deuxième contient les salaires et le troisième les indices des prix à la consommation.

1.4 Calcul de l'IPC

En ajustant les salaires des basketteurs entre 1996 et 2017 nous visons à eliminer l'effet de l'inflation sur les prix. Cela nous permettra de nous concentrer uniquement sur les performances des joueurs en tant que variable explicatives. En utilisant l'IPC, on garantit que les revenus des joueurs restent adaptés à l'inflation

2 Statistique descriptive

2.1 Analyse des données

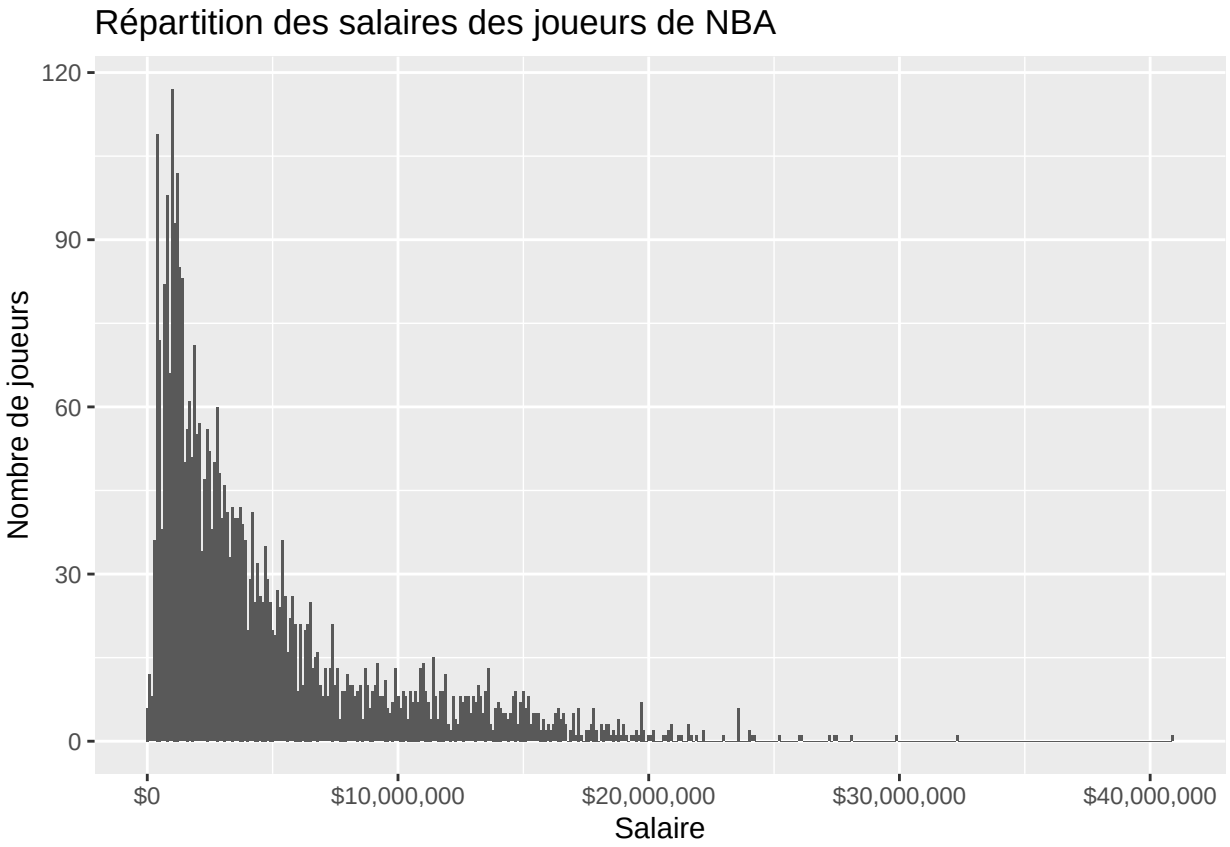
```
##   Season_Start         Player_Name        Team           age
##  2017   : 206   Kobe Bryant  :  19   UTA    : 168   Min.   :19.00
##  2011   : 201   Dirk Nowitzki:  17   WAS    : 160   1st Qu.:24.00
##  2016   : 201   Kevin Garnett:  15   SAS    : 159   Median :27.00
##  2015   : 200   Tim Duncan   :  14   CLE    : 154   Mean   :27.46
##  2002   : 196   Tony Parker  :  14   LAL    : 154   3rd Qu.:30.00
##  2006   : 193   Manu Ginobili:  13   GSW    : 151   Max.   :42.00
##  (Other):2516   (Other)      :3621   (Other):2767
##  player_height   player_weight              college          country
##  Min.   :160.0   Min.   : 60.33   None           : 624   USA      :3056
##  1st Qu.:195.6   1st Qu.: 92.99   North Carolina: 127   France   :  61
##  Median :203.2   Median :102.06   Duke           : 119   Canada   :  50
##  Mean   :201.7   Mean   :101.86   Arizona        : 109   Spain    :  42
##  3rd Qu.:208.3   3rd Qu.:111.13   Kentucky       : 104   Brazil   :  35
##  Max.   :231.1   Max.   :140.61   UCLA           : 100   Lithuania:  30
##                                   (Other)        :2530   (Other)  : 439
##      draft_year      draft_number        gp              pts
##  Undrafted: 347   Undrafted: 348   Min.   : 1.00   Min.   : 0.00
##  2003     : 199   1        : 161   1st Qu.:50.00   1st Qu.: 5.30
##  2001     : 185   5        : 139   Median :68.00   Median : 9.30
##  1998     : 176   4        : 138   Mean   :61.05   Mean   :10.37
##  2005     : 162   3        : 136   3rd Qu.:78.00   3rd Qu.:14.40
##  2004     : 161   7        : 125   Max.   :82.00   Max.   :35.40
##  (Other)  :2483   (Other)  :2666
##       reb              ast           net_rating          oreb_pct
##  Min.   : 0.000   Min.   : 0.000   Min.   :-88.3000   Min.   :0.00000
##  1st Qu.: 2.400   1st Qu.: 0.800   1st Qu.: -4.2000   1st Qu.:0.02400
##  Median : 3.800   Median : 1.600   Median :  0.3000   Median :0.04700
##  Mean   : 4.446   Mean   : 2.237   Mean   : -0.1532   Mean   :0.05844
##  3rd Qu.: 5.900   3rd Qu.: 3.000   3rd Qu.:  4.4000   3rd Qu.:0.08900
##  Max.   :16.100   Max.   :11.700   Max.   : 55.3000   Max.   :0.36800
##
##     dreb_pct         usg_pct           ts_pct         ast_pct
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.1010   1st Qu.:0.1550   1st Qu.:0.496   1st Qu.:0.0700
##  Median :0.1400   Median :0.1880   Median :0.531   Median :0.1100
##  Mean   :0.1492   Mean   :0.1933   Mean   :0.525   Mean   :0.1378
##  3rd Qu.:0.1900   3rd Qu.:0.2280   3rd Qu.:0.563   3rd Qu.:0.1840
##  Max.   :0.3890   Max.   :0.5000   Max.   :1.000   Max.   :1.0000
##
##      Salary           log_Salary     USA
##  Min.   :    8012   Min.   : 8.989   0: 657
##  1st Qu.: 1355370   1st Qu.:14.120   1:3056
##  Median : 3096681   Median :14.946
##  Mean   : 4811055   Mean   :14.897
##  3rd Qu.: 6392428   3rd Qu.:15.671
##  Max.   :40943855   Max.   :17.528
##
```

La présentation des données sur les basketteurs de 1996 à 2017 révèle plusieurs points clés. En termes de

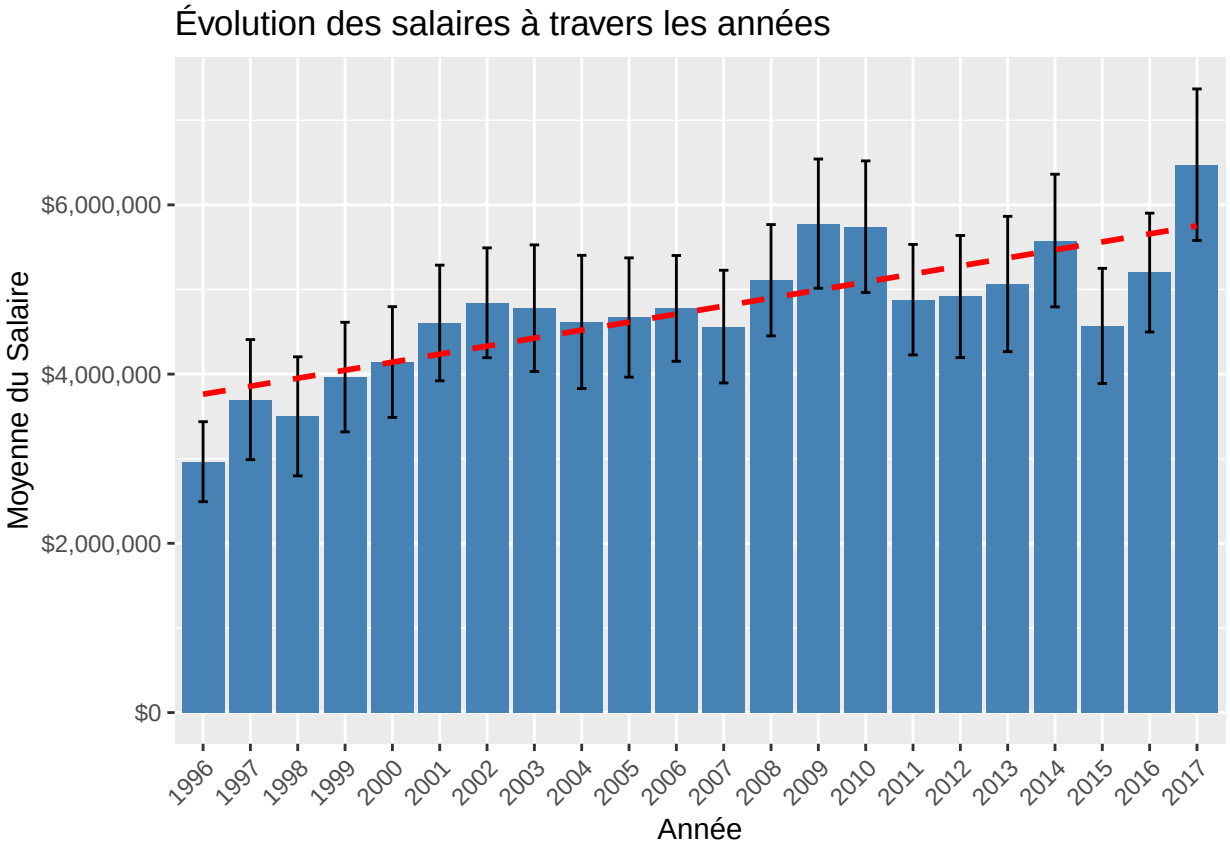
démographie, des joueurs tels que Kobe Bryant, Dirk Nowitzki, et Vince Carter se démarquent par leur longévité avec 20, 18, et 18 saisons respectivement. Les États-Unis dominent la nationalité des joueurs, représentant la grande majorité, suivi par la France et le Canada. En ce qui concerne les statistiques de jeu, la moyenne de points (pts) est d'environ 8,5, avec une médiane de 7,1, tandis que la moyenne de salaire est d'environ 4,4 millions de dollars, avec une médiane de 2,6 millions de dollars. Ces données offrent un aperà§u de la diversité des joueurs et de la variabilité des performances et des rémunérations au fil des années.

On constate que le nombre maximal de match joué par un joueur est de 85 or une saison de NBA compte 82 matchs. Cela s'explique par le fait que certains joueurs ont été transférés en cours de saison.

2.2 Analyse de la variable Salaire

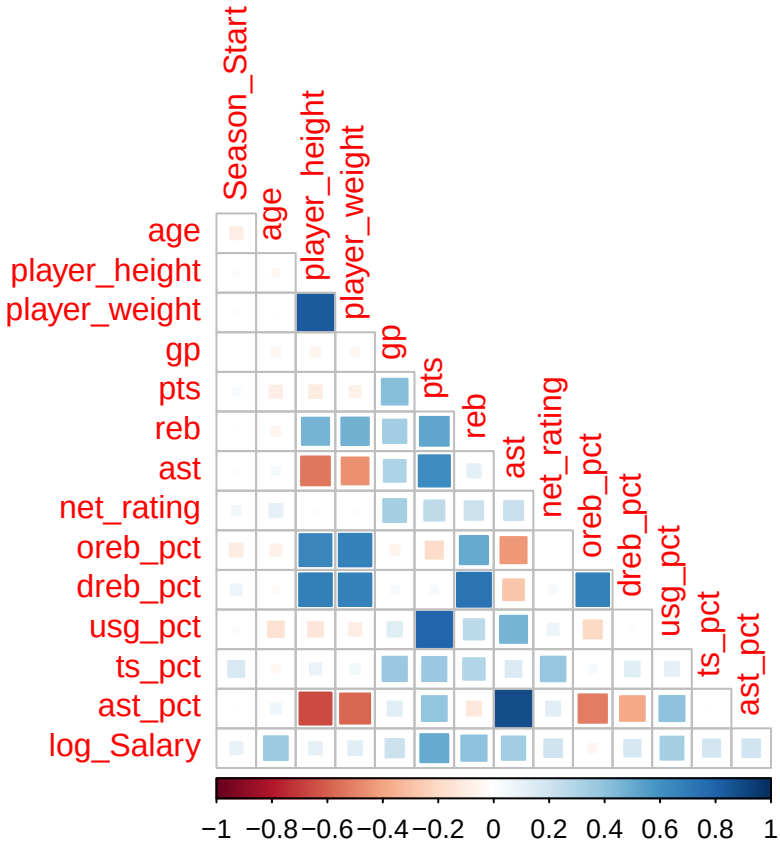


On observe que la distribution des salaires est asymétrique à droite. La majorité des joueurs gagnent entre 0 et 5 millions de dollars. On observe une queue de distribution à droite, avec des salaires allant jusqu'à 40 millions de dollars. Cela suggère que les salaires des joueurs sont très variables, avec une grande disparité entre les joueurs les mieux payés et les joueurs les moins bien payés. Les joueurs payés sont souvent ceux qui ont les meilleurs performances sur le terrain. On peut donc s'attendre à ce que les joueurs les mieux payés soient ceux qui ont les meilleures statistiques de jeu. Pour vérifier cette hypothèse, on peut examiner la relation entre le salaire et les statistiques de jeu.



On observe une tendance à la hausse des salaires au fil des années. Cela s'explique par le fait que les revenus de la NBA ont augmenté au fil des années. En effet, les revenus de la NBA sont passés de plus de 2 millions de dollars en 1996 à 7 millions de dollars en 2017. Cette augmentation des revenus est due à l'augmentation des droits de diffusion, des droits de parrainage et des droits de billetterie. Cette augmentation des revenus a permis aux équipes de la NBA d'augmenter les salaires des joueurs afin de rester compétitifs .

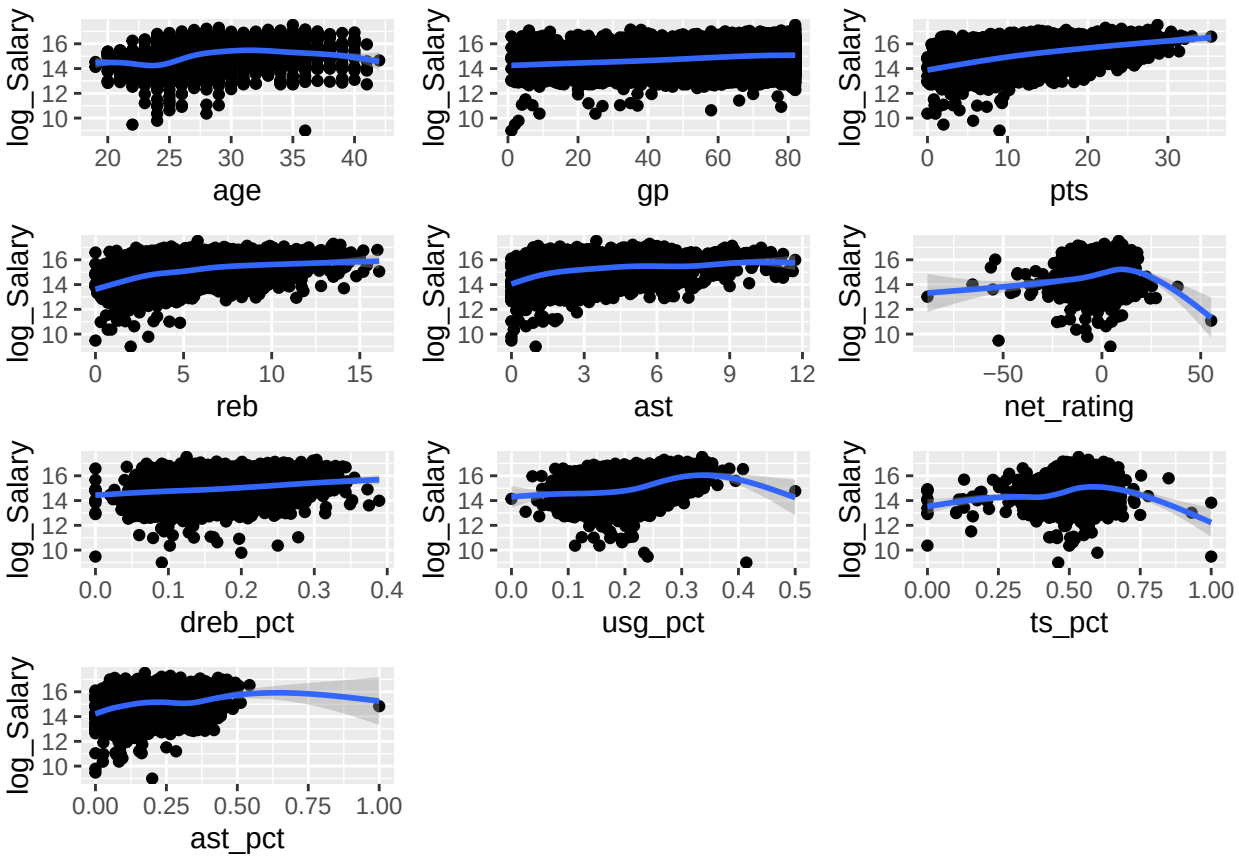
2.3 Analyse de corrélation entre nos variables



Ce graphique nous indique que le salaire est corrélé positivement avec l'age , les points, les rebonds, les passes décisives, le pourcentage de rebonds offensifs, le pourcentage de rebonds défensifs, le pourcentage d'utilisation ainsi que le pourcentage de tirs réussis. Cela confirme que les joueurs avec de meilleurs statistiques de jeu sont mieux payés.

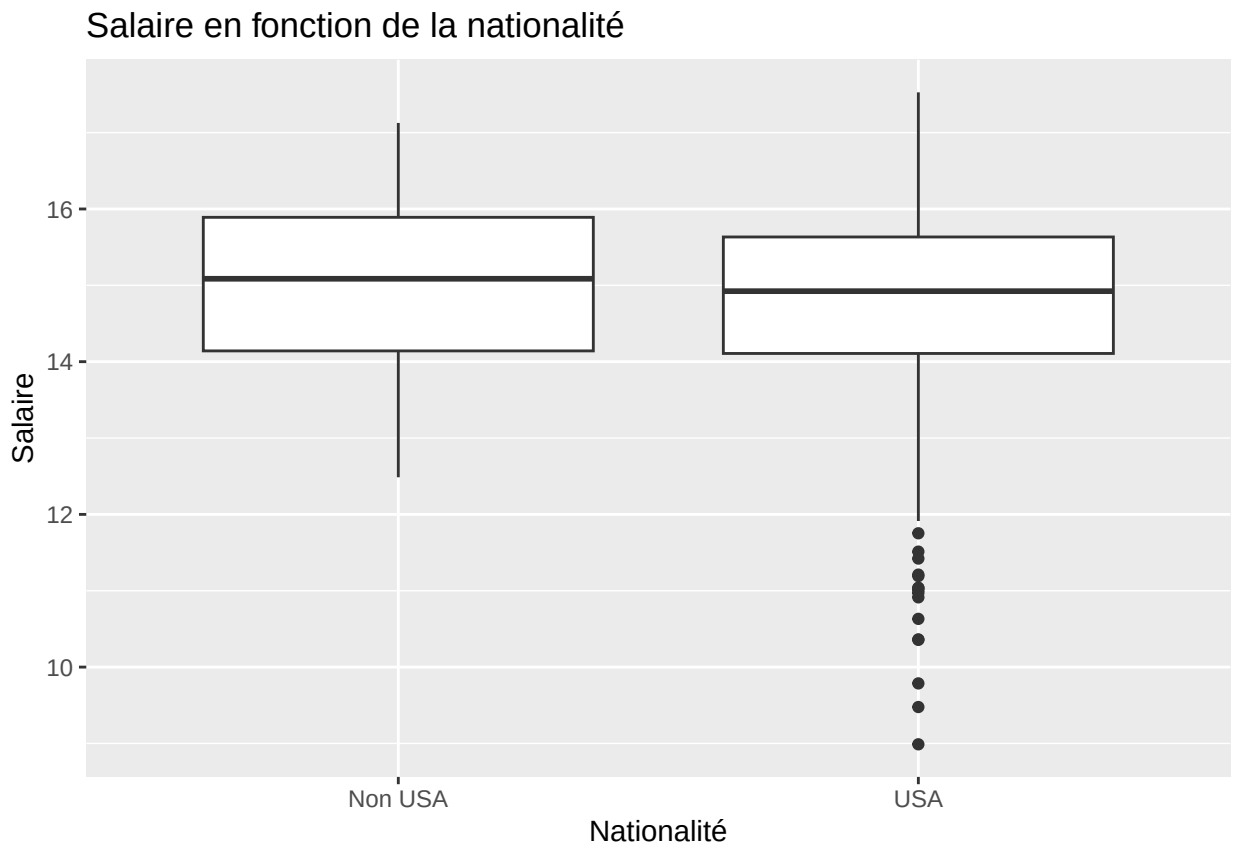
On constate également que certaine variable explicative sont corrélées entre elles. C'est le cas par exemple des points et des passes décisives. On peut donc s'attendre à ce que ces deux variables explicatives aient un impact similaire sur le salaire. Il serait donc interessant a l'avenir de faire une analyse en composante principale afin de réduire le nombre de variable explicative et de supprimer les variables explicatives corrélées entre elles.

Pour vérifier nos hypothèses, on peut examiner la relation entre le salaire et les statistiques de jeu.



On peut voir que les variables sont très différentes les unes des autres. Certaines variables sont continues, d'autres sont discrètes. On constate sur les graphiques que certains points semblent aberrants. Par exemple, on peut voir que le joueur qui receptionne le plus de rebond défensifs. On peut aussi voir que certains joueurs ont un poids très élevé, ce qui peut être du à leur grande taille.

On observe que le pourcentage maximal de certaines variables, telles que dreb_pct, ast_pct, oreb_pct et usg_pct, atteint 100%. Cette occurrence pourrait potentiellement résulter du fait que certains joueurs ont joué un nombre limité de matchs au cours de leur carrière, ce qui aurait réduit leurs occasions et, par conséquent, minimisé leurs chances d'échec. On préfère ici les supprimer car ses valeurs releves plus de l'exceptionnel.



Cela nous indique que les joueurs non américains sont mieux payés que les joueurs américains. Cela peut s'expliquer par le fait que les joueurs non américains sont souvent des joueurs de renommée mondiale, qui ont donc un salaire plus élevé. On peut aussi penser que les joueurs américains sont plus nombreux que les joueurs non américains, et que les joueurs américains sont donc moins bien payés car ils sont plus nombreux. On peut donc s'attendre à ce que la nationalité ait un impact sur le salaire des joueurs.

3 Régression linéaire

La régression linéaire en économétrie, lorsqu'elle est appliquée au contexte du basket, peut permettre d'explorer la relation entre les salaires des joueurs, qui peut être influencée par diverses variables explicatives telles que les performances, l'expérience, ou "l'image de marque" qu'un joueur renvoie via sa notoriété. En se concentrant sur les performances individuelles, cette approche statistique permet de quantifier de manière précise l'influence de ces facteurs sur les rémunérations, offrant ainsi des insights cruciaux pour les décisions de gestion, les négociations contractuelles et l'optimisation des ressources financières dans le monde professionnel du basket.

```
##
## Call:
## lm(formula = log_Salary ~ age + gp + pts + reb + ast + net_rating +
##     dreb_pct + usg_pct + ts_pct + ast_pct + Season_Start, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7647 -0.4743  0.0840  0.5286  2.0887
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.945595   0.178276  61.397  < 2e-16 ***
## age           0.106598   0.003163  33.705  < 2e-16 ***
## gp           -0.001288   0.000748  -1.721  0.08525 .
## pts           0.053366   0.006723   7.938 2.69e-15 ***
## reb           0.075116   0.011927   6.298 3.37e-10 ***
## ast           0.142015   0.022166   6.407 1.67e-10 ***
## net_rating    0.002172   0.001960   1.108  0.26785
## dreb_pct      0.531467   0.451088   1.178  0.23880
## usg_pct       1.018704   0.540034   1.886  0.05932 .
## ts_pct       -0.652460   0.230936  -2.825  0.00475 **
## ast_pct      -2.064708   0.418565  -4.933 8.46e-07 ***
## Season_Start  0.020637   0.002067   9.986  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7706 on 3698 degrees of freedom
## Multiple R-squared:  0.4722, Adjusted R-squared:  0.4707
## F-statistic: 300.8 on 11 and 3698 DF,  p-value: < 2.2e-16
```

On constate que la majorités des variables ont des effets significatifs individuellement grâce au test de student, mais l'ensemble du modèle n'explique qu'une proportion limitée de la variabilité des données, 52% . Cela peut être dû à des interactions complexes entre les variables, à des colinéarités ou à d'autres problèmes dans la spécification du modèle qu'il va falloir corriger.

Les variables net_rating, usg_pct et ts_pct n'ont pas d'influences significative sur le salaire. Pourtant, si on régresse le salaire des joueurs sur ces variables là et une constante à chaque fois, voici ce que l'on observe :

```
##
## Call:
## lm(formula = Salary ~ net_rating, data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -8833840 -3176084 -1522310  1624430 34997961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4834425      76779   62.97   <2e-16 ***
## net_rating    130761      10400   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4676000 on 3708 degrees of freedom
## Multiple R-squared:  0.04089,    Adjusted R-squared:  0.04063
## F-statistic: 158.1 on 1 and 3708 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = Salary ~ usg_pct, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12453799  -3063132  -1238523   1979091  31184317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1881023     270363  -6.957 4.08e-12 ***
## usg_pct     34644527    1348313  25.695  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4399000 on 3708 degrees of freedom
## Multiple R-squared:  0.1511, Adjusted R-squared:  0.1509
## F-statistic: 660.2 on 1 and 3708 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = Salary ~ ts_pct, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10831980  -3236813  -1581042   1671694  36032725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1861400     606383   -3.07  0.00216 **
## ts_pct      12706436    1144950   11.10  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4697000 on 3708 degrees of freedom
## Multiple R-squared:  0.03215,    Adjusted R-squared:  0.03189
## F-statistic: 123.2 on 1 and 3708 DF,  p-value: < 2.2e-16
```

Si on effectue une régression linéaire simple du salaire des joueurs sur chacune des autres variables du data frame, on peut remarquer que dans tous les cas, le test de nullité du coefficient est rejeté à 1%.

On peut donc en conclure que ces variables ont un effet significatif sur le salaire des joueurs mais que cet effet est négligeable par rapport aux autres variables. On peut donc les supprimer du modèle.

```
##
## Call:
## lm(formula = log_Salary ~ age + pts + reb + ast + dreb_pct +
##     Season_Start, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6215 -0.4874  0.0776  0.5346  2.0755
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.543060   0.102736 102.623  < 2e-16 ***
## age           0.106010   0.003105  34.137  < 2e-16 ***
## pts           0.063871   0.003471  18.403  < 2e-16 ***
## reb           0.071665   0.010445   6.861 7.95e-12 ***
## ast           0.043961   0.008925   4.925 8.79e-07 ***
## dreb_pct      0.855717   0.415228   2.061   0.0394 *
## Season_Start  0.019564   0.002043   9.577  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7739 on 3703 degrees of freedom
## Multiple R-squared:  0.467,  Adjusted R-squared:  0.4661
## F-statistic: 540.6 on 6 and 3703 DF,  p-value: < 2.2e-16
```

On constate que le coefficient de détermination est légèrement plus faible mais que toutes les variables sont significatives pour un niveau de confiance à 99% .

Nous pourrons également analyser l'effet croisées entre le nombre de point marqué et la nationalité des joueurs pour voir si les joueurs étranger avec un bon score sont mieux rémunéré

```
##
## Call:
## lm(formula = log_Salary ~ age + pts:USA + pts + USA + reb + ast +
##     dreb_pct + Season_Start, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6095 -0.4863  0.0807  0.5374  2.0939
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.713263   0.117947  90.831  < 2e-16 ***
## age           0.106574   0.003111  34.262  < 2e-16 ***
## pts           0.053429   0.006024   8.870  < 2e-16 ***
## USA1         -0.194242   0.067537  -2.876  0.00405 **
## reb           0.073016   0.010478   6.968 3.78e-12 ***
## ast           0.042812   0.008930   4.794 1.70e-06 ***
## dreb_pct      0.743685   0.416612   1.785  0.07433 .
## Season_Start  0.018797   0.002071   9.076  < 2e-16 ***
## pts:USA1      0.012276   0.005660   2.169  0.03014 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7732 on 3701 degrees of freedom
## Multiple R-squared:  0.4682, Adjusted R-squared:  0.467
## F-statistic: 407.3 on 8 and 3701 DF,  p-value: < 2.2e-16
```

On remarque qu'il existe effectivement une interaction entre le nombre de point marqué et la nationalité des joueurs. En effet, le coefficient de la variable USA1 est significatif et négatif. On peut donc en conclure que les joueurs étrangers sont généralement mieu rémunéré que les joueurs américains. Cependant, le coefficient de la variable pts:USA1 est positif et significatif. On peut donc en conclure que cette écart a tendance a diminué par rapport au nombre de point marqués.

```
## Analysis of Variance Table
##
## Model 1: log_Salary ~ age + pts + reb + ast + dreb_pct + Season_Start
## Model 2: log_Salary ~ age + pts:USA + pts + USA + reb + ast + dreb_pct +
##     Season_Start
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   3703 2217.9
## 2   3701 2212.7  2    5.1649 4.3193 0.01338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

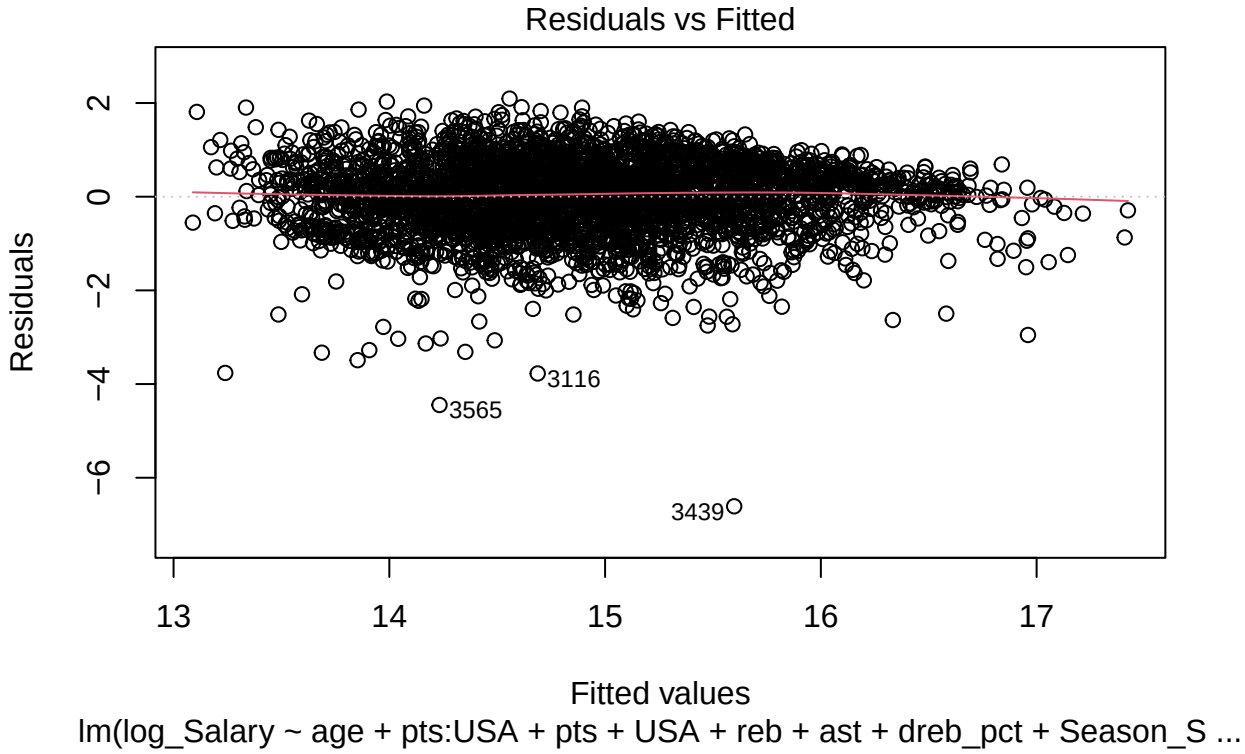
On remarque que le test de Fisher est significatif. On peut donc en conclure que le modèle avec l'interaction entre le nombre de point marqué et la nationalité des joueurs est meilleur que le modèle sans cette interaction.

4 Estimation robuste

L'objectif de cette partie est diagonostiquer les problèmes de notre modèle et de les corriger. Nous allons donc effectuer une estimation robuste pour corriger les problèmes d'hétéroscédasticité, d'autocorrélation et de non normalité des résidus.

4.1 Analyse des résidus

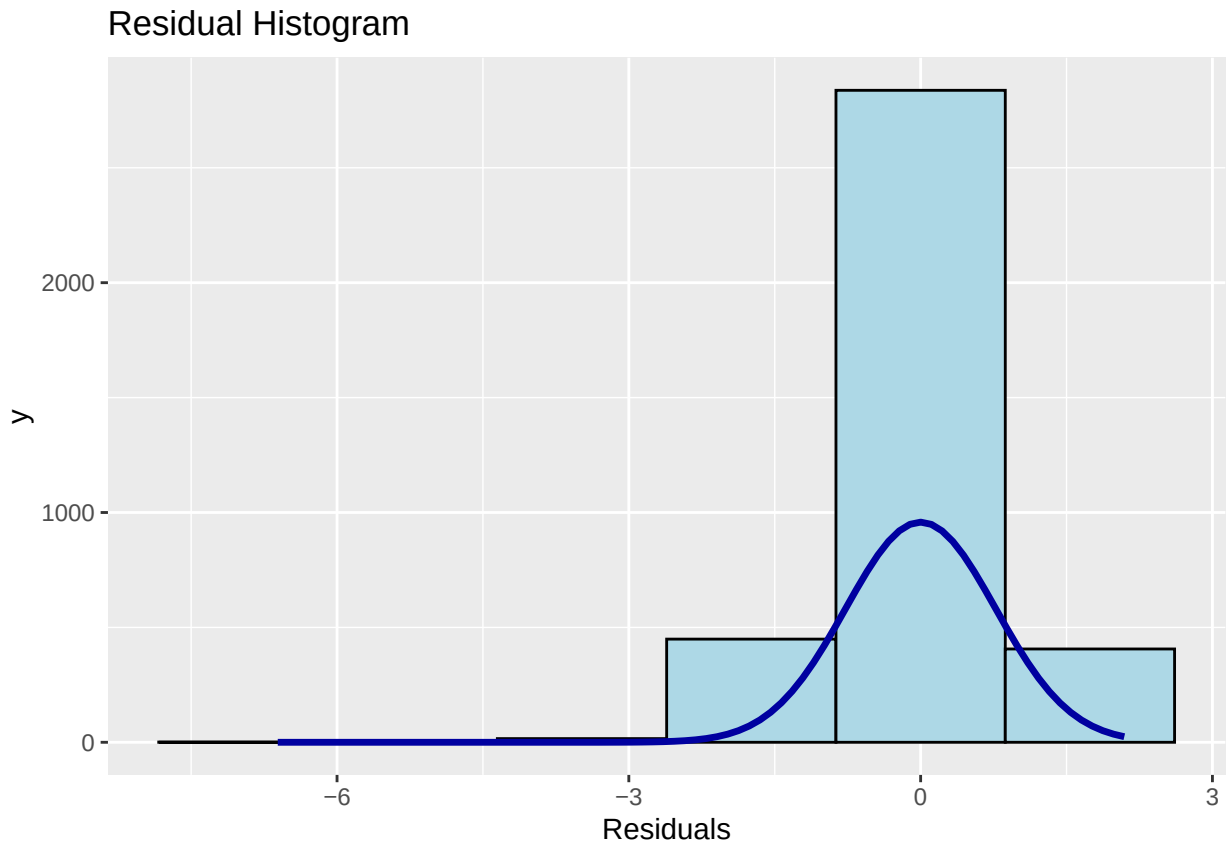
4.1.1 représentations graphiques des résidus

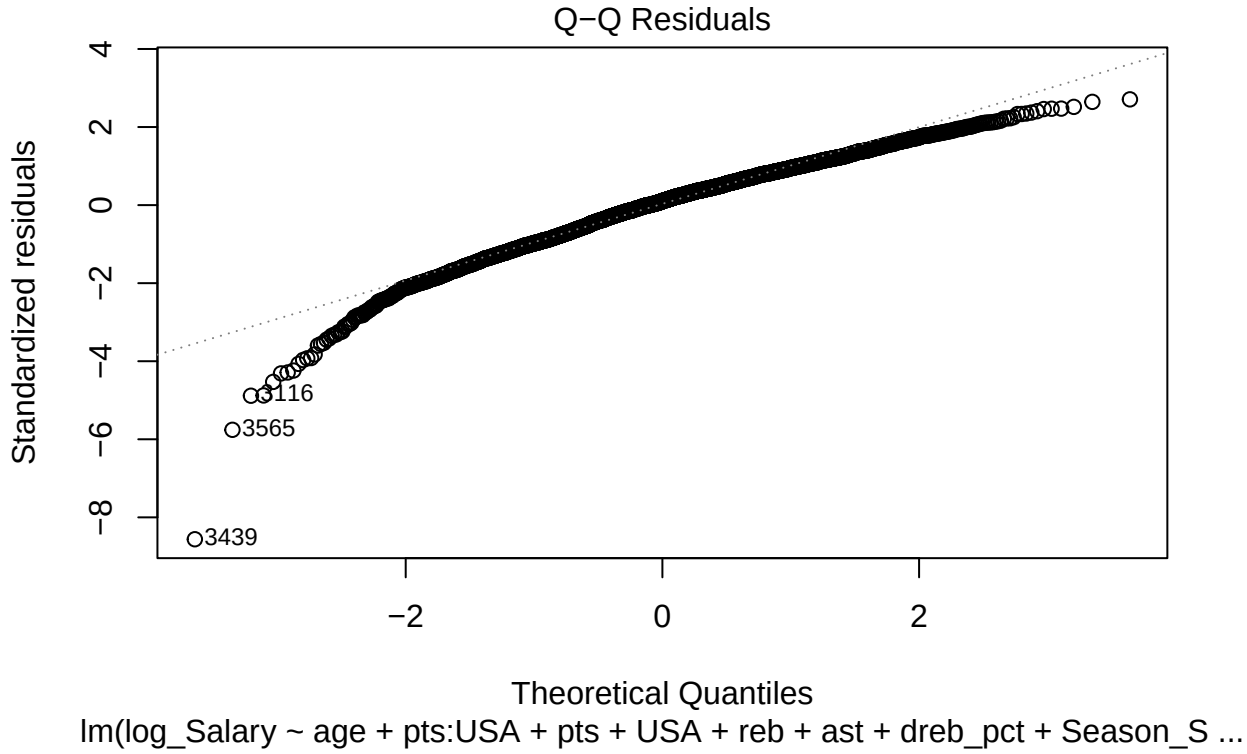


On remarque que la variance des résidus n'est pas constante. Il est possible que l'hypothèse d'homoscédasticité ne soit pas vérifiée. De plus on constate que les résidus ne sont pas centrés autour de 0. Il est donc possible que l'hypothèse de normalité ne soit pas vérifiée.Nous allons donc effectuer une estimation robuste pour corriger ces problèmes.

4.1.2 Test sur les résidus

Nous nous poserons ici la question de la normalité de nos résidus ainsi qu'au test de l'hypothèse de nullité de la moyenne des résidus pour vérifier si notre modèle est bien ajusté aux données.


Residual Histogram
2000
1000
0
y
−6
−3
0
3
Residuals




```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.97214, p-value < 2.2e-16
```

Bien que le test de Shapiro-Wilk indique une significative déviation de la normalité pour les résidus de notre modèle, il est important de noter que cet échantillon est de taille importante. De plus, la représentation graphique des résidus suggère une similarité avec une distribution normale. Ces résultats pourraient indiquer que, malgré la détection statistique de la non-normalité, la robustesse des tests de régression linéaire est préservée dans ce contexte de grand échantillon.

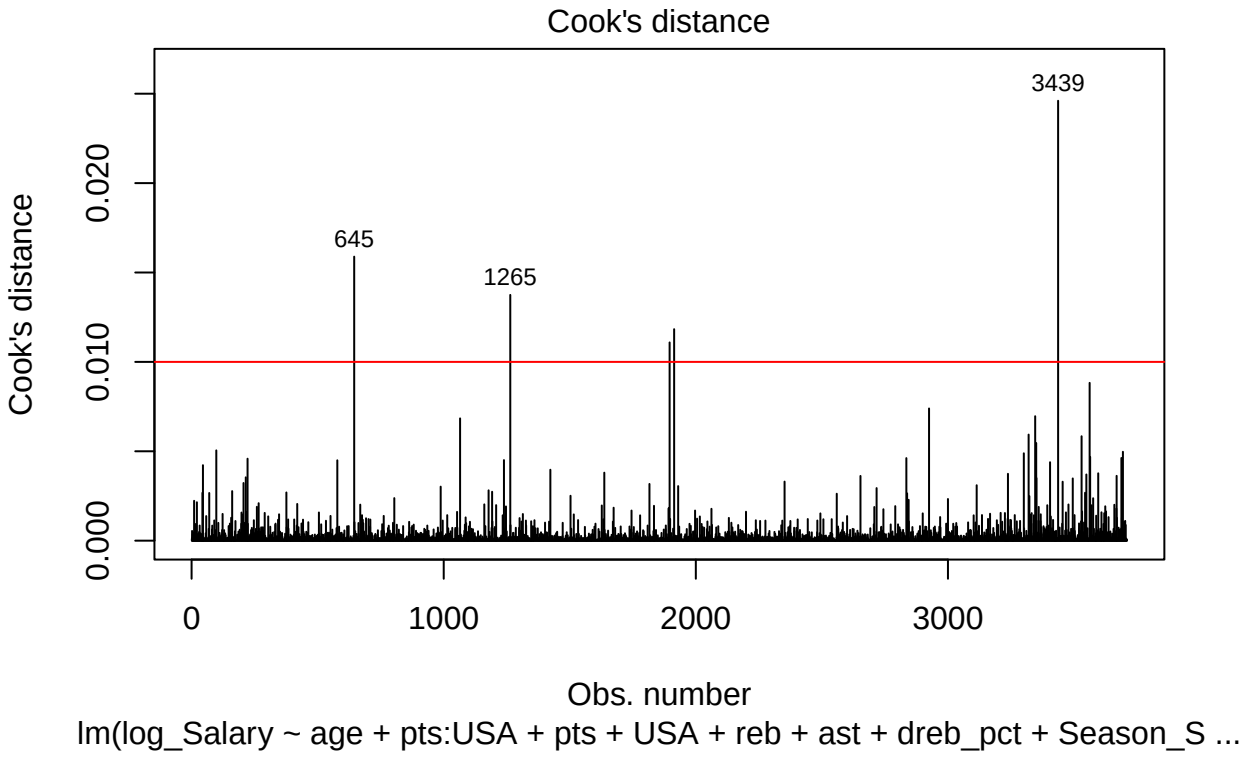
Nous pouvons alors tout de même effectuer un test

```
##
##  One Sample t-test
##
## data:  reg$residuals
## t = 2.9758e-15, df = 3709, p-value = 0.5
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.02086344         Inf
## sample estimates:
##    mean of x
## 3.773553e-17
```

D'après le test de Student, On conclut que la moyenne des résidus est significativement supérieur à 0. On peut donc en conclure que le modèle n'est pas parfaitement ajusté aux données. Cela suggère une tendance à sous-estimer les résultats, ce qui peut être dû à des variables manquante dans notre modèle n'ayant pas de lien avec les performances des joueurs, il existe donc d'autre facteur qui explique le salaire des joueurs indépendamment de leurs statistiques de jeu tel que la rentabilité commerciales d'un joueurs pour son équipe par exemple.

4.1.3 Identification des points influents et suppression des points leviers

Nous pouvons a présent analyser les points influents et les points leviers de notre modèle pour voir si certaines observations ont un effet pouvant influencer de facon biaisé nos estimations.



Après avoir detecter les points leviers de notre modèle nous allons les supprimer pour voir si cela améliore la significativité de notre modèle.

```
##
## Call:
## lm(formula = log_Salary ~ age + pts:USA + pts + USA + reb + ast +
##     dreb_pct + Season_Start, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4581 -0.4880  0.0753  0.5329  2.0776
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

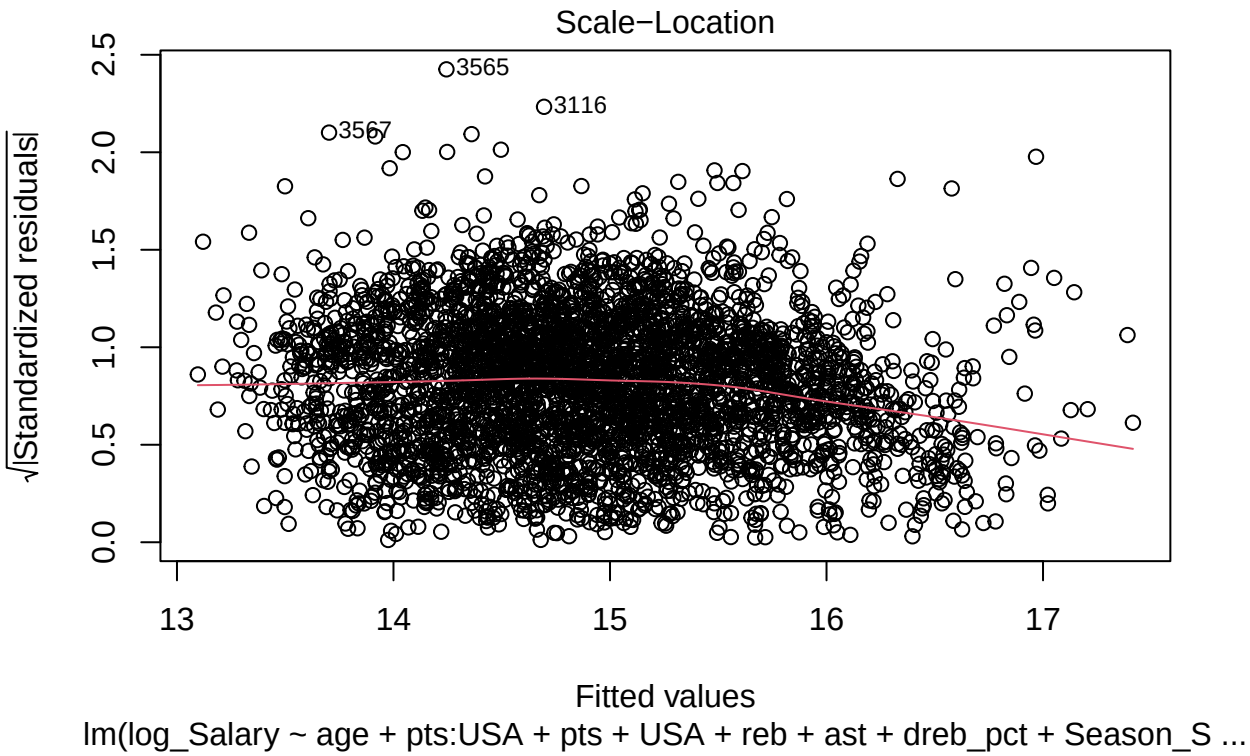
```
## (Intercept)  10.694430   0.116019  92.178  < 2e-16 ***
## age           0.107303   0.003056  35.112  < 2e-16 ***
## pts           0.054298   0.005921   9.171  < 2e-16 ***
## USA1         -0.182733   0.066326  -2.755   0.0059 **
## reb           0.071156   0.010389   6.849 8.65e-12 ***
## ast           0.040606   0.008770   4.630 3.78e-06 ***
## dreb_pct      0.743933   0.414597   1.794   0.0728 .
## Season_Start  0.019082   0.002035   9.378  < 2e-16 ***
## pts:USA1      0.011588   0.005557   2.085   0.0371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7589 on 3696 degrees of freedom
## Multiple R-squared:  0.4757, Adjusted R-squared:  0.4746
## F-statistic: 419.2 on 8 and 3696 DF,  p-value: < 2.2e-16
```

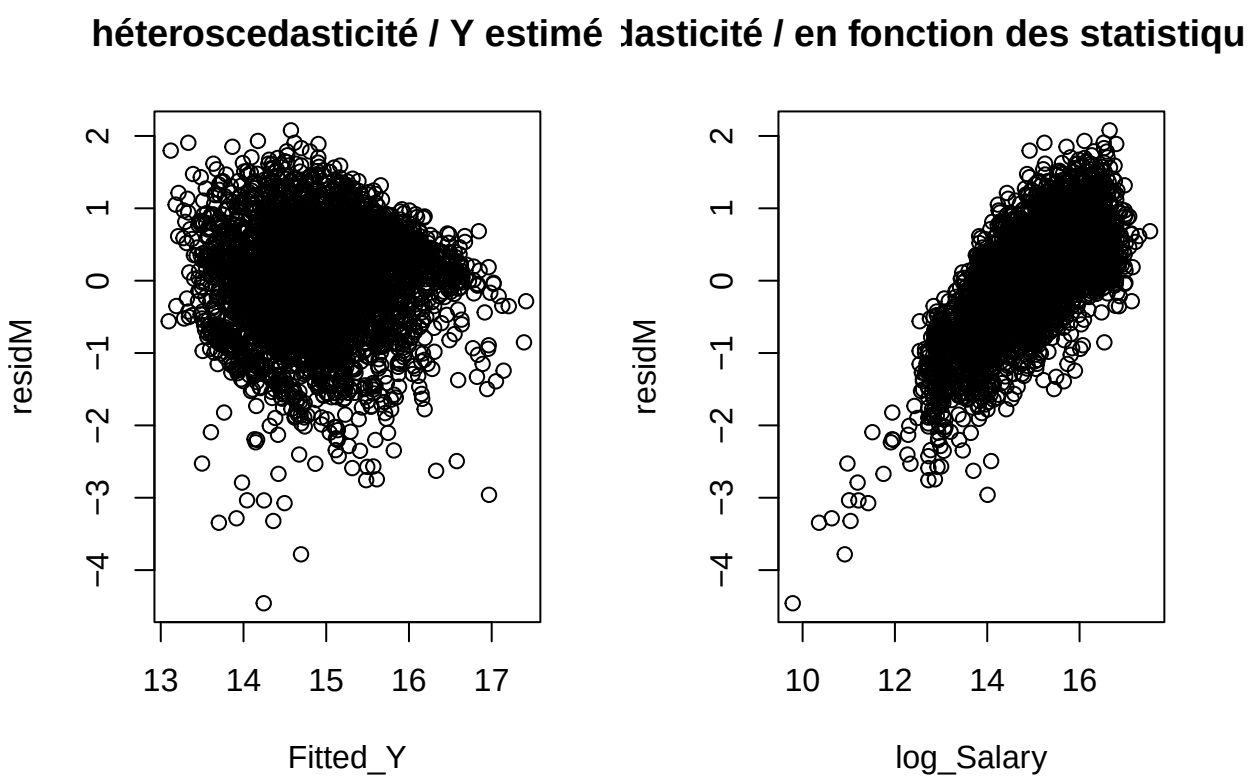
En ayant supprimer les d'influences de notre modèle nous avons augmenter la significativité de celui ci. En effet, la statistique de Fisher est plus élevé et le coefficient de détermination est plus important. On peut donc en conclure que notre modèle est plus robuste.

4.2 Detection d'anomalies

4.2.1 Hétéroscédasticité

L'hétéroscédasticité en régression linéaire se manifeste lorsque la variance des résidus n'est pas constante à travers les niveaux de la variable prédite. Cela peut compromettre la validité des tests statistiques et des intervalles de confiance. Afin de remédier à ce problème, des méthodes comme la transformation des variables ou l'utilisation de modèles robustes peuvent être envisagées.


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3565
3116
3567
13
14
15
16
17
Fitted values
lm(log_Salary ~ age + pts:USA + pts + USA + reb + ast + dreb_pct + Season_S ...



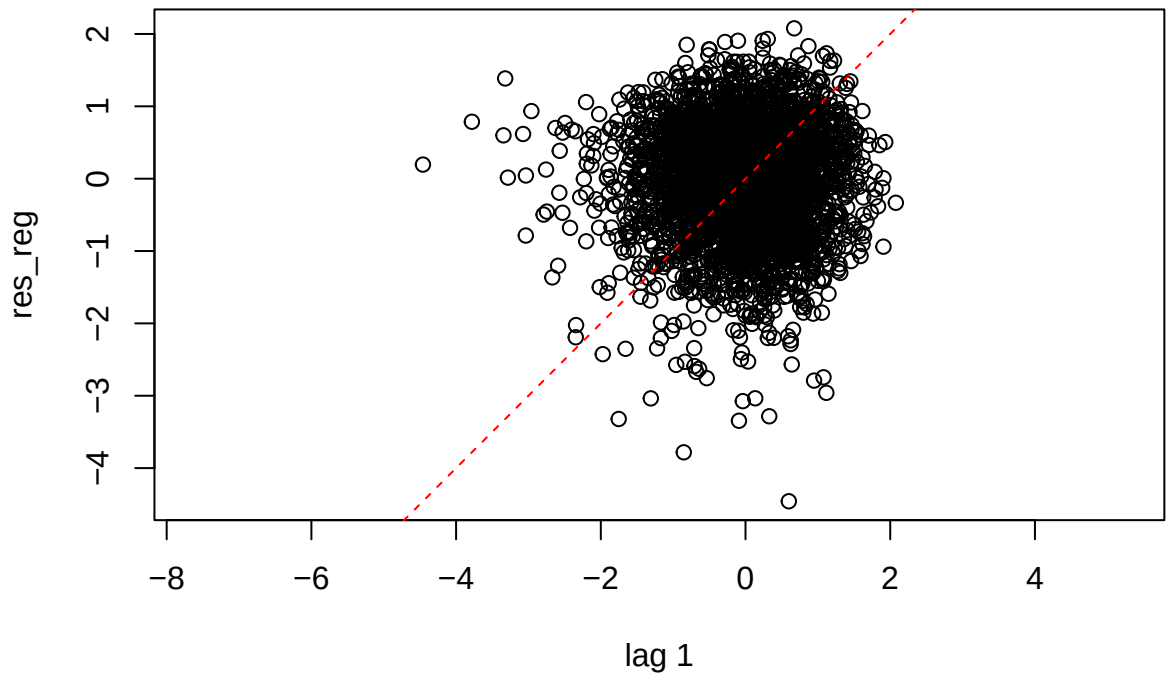

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 69.605, df = 8, p-value = 5.889e-12
```

Sur le premier graphique on peut constater que la ligner rouge n'est pas exactement horizontale, ce qui indique une hétéroscédasticité de nos résidus. On constate que lla ligne a une pente négative, ce qui indique une diminution de la variance avec les valeurs de X.

Sur le graphique des résidus en fonction des variables explicatives, on peut voir que la variance des résidus augmente avec la valeur de la variable explicative. On peut donc en conclure que notre modèle souffre d'hétéroscédasticité.

4.2.2 Auto-corrélation

En régression linéaire, l'autocorrélation se produit lorsque les résidus du modèle, représentant les erreurs de prédiction, présentent une corrélation systématique entre eux. Cela peut fausser les résultats et nécessite une correction pour garantir la validité des analyses statistiques.




```
##
##  Durbin-Watson test
##
## data:  reg
## DW = 1.9303, p-value = 0.01627
## alternative hypothesis: true autocorrelation is greater than 0
```

Le graphique révèle une relation particulière entre les résidus en t et t-1, on semble voir une autocorrelation positive des aléas. Or le test de Durbin Watson nous indique que l'autocorrelation est négative. Cela peut être du à la présence de variable manquante dans notre modèle qui explique le salaire des joueurs indépendamment de leurs statistiques de jeu tel que la rentabilité commerciales d'un joueurs pour son équipe par exemple.

4.2.3 Correction de l'hétéroscédasticité et de l'autocorrélation

Pour avoir un meilleur modèle, nous devons donc corriger l'hétéroscédasticité et l'auto-corrélation. Pour cela nous allons utiliser la méthode de White qui consiste à calculer la matrice de variance covariance des paramètres. Cette matrice ne souffre ni d'hétéroscédasticité ni d'autocorrélation.

```
##
## Comparaison de l'estimation MCO et MCO correction WHITE
## ================================================================
##                       Variable dépendante : 'CO2 par habitant'
##                       ----------------------------------------
##                                   MCO                MCOWhite
```

```
## ---------------------------------------------------------------
## age                              0.107***             0.107***
##                                  (0.003)              (0.003)
##
## pts                              0.054***             0.054***
##                                  (0.006)              (0.006)
##
## USA1                            -0.183***            -0.183**
##                                  (0.066)              (0.071)
##
## reb                              0.071***             0.071***
##                                  (0.010)              (0.011)
##
## ast                              0.041***             0.041***
##                                  (0.009)              (0.008)
##
## dreb_pct                          0.744*               0.744*
##                                  (0.415)              (0.442)
##
## Season_Start                     0.019***             0.019***
##                                  (0.002)              (0.002)
##
## pts:USA1                         0.012**              0.012**
##                                  (0.006)              (0.005)
##
## Constant                        10.694***            10.694***
##                                  (0.116)              (0.117)
##
## ---------------------------------------------------------------
## Observations                      3,705
## R2                                0.476
## Adjusted R2                       0.475
## Residual Std. Error      0.759 (df = 3696)
## F Statistic          419.151*** (df = 8; 3696)
## ===============================================================
## Note:                               *p<0.1; **p<0.05; ***p<0.01
```

Nous obtenons ici notre tableau de comparaison de nos deux modèles. On peut voir que les paramètres sont sans biais et que la méthode de White permet de recalculer la variance des paramètres. Certaines variables sont donc moins ou plus significatives. Mais la plus part le sont plus, seulement les variables ast et pts:USA sont moins significatives. Nous allons donc garder ce modèle qui est plus robuste que le précédent.

Conclusion

Afin de répondre à notre problématique, nous avons réalisé une étude sur les salaires des joueurs de NBA dans le but de déterminer la part du salaire pouvent être imputé au performance individuelle de chaque sportifs. Nous avons donc étudié les facteurs qui influencent le salaire des joueurs. Pour cela, on a réalisé une régression linéaire multiple et nous avons pu déterminer les variables explicatives les plus significatives et les plus influentes sur le salaire des joueurs :

```
##
## ===============================================================
```

```
##                                  Variable dépendante : 'Salaire'
##                                  -------------------------------
##                                              MCO
## ------------------------------------------------------------------
## Âge                                       0.107***
##                                            (0.003)
##
## Points Marqués                            0.054***
##                                            (0.006)
##
## Joueur Américain                          -0.183**
##                                            (0.071)
##
## Nombre de Rebonds                         0.071***
##                                            (0.011)
##
## Nombre de Passes Décisives                0.041***
##                                            (0.008)
##
## Pourcentage de Rebonds Défensifs           0.744*
##                                            (0.442)
##
## Année de la saison                        0.019***
##                                            (0.002)
##
## Points Américain                           0.012**
##                                            (0.005)
##
## Constante                                 10.694***
##                                            (0.117)
##
## ==================================================================
## ==================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

A travers ce modèle nous avons pu determiner que les performances des joueurs peuvent expliquer 47.5% de la variance des salaires.

Les résultats de notre étude indiquent de manière concluante que les performances individuelles des joueurs jouent un rôle crucial dans la détermination de leur rémunération. Les performances antérieures sur le terrain se révèlent être un facteur déterminant dans la négociation des contrats et la fixation des salaires des joueurs de la NBA. Ces constatations soulignent l'importance pour les joueurs de maintenir des niveaux élevés de performance pour garantir une compensation financière correspondante.

Cependant, il est essentiel de noter que les performances individuelles ne sont pas les seuls facteurs influençant les salaires des joueurs. Notre étude a révélé que la nationalité des joueurs, l'année de la saison, l'âge et l'origine du joueur sont également des éléments influençant la rémunération. En effet, la diversité de ces facteurs sont des variables importantes à considérer dans la compréhension complète des déterminants salariaux dans le contexte de la NBA. Par ailleurs, il est important de souligner que le salaire d'un joueur est également influencé par son image publique et la notoriété qu'il génère. La visibilité médiatique, les partenariats commerciaux la perception publique ainsi que sa popularité peuvent constituer des éléments déterminants dans la négociation de contrats et ainsi exercer une influence significative sur les opportunités financières qui lui sont offertes en plus de ses performances sportives. Cette dimension supplémentaire renforce l'idée que la dynamique des salaires dans le domaine du basketball professionnel est complexe et multifactorielle, combinant des aspects sportifs et des considérations plus larges liées à l'image et à la

notoriété.